

Sketch Engine

- Home
- Register
- Log in
- Lost password

Support

- User guide
- Contact support

Log in

[Authenticate using a single sign on \(SSO\) service](#) ?

User name

Password

Lost your password? Please [click here](#).

Don't have an account? Please [register](#).

User name

Title

First name

Last name

E-mail

If you do not receive your password within a few minutes, check your spam folder and contact support@sketchengine.co.uk.

Country

Site licence key

Terms of use

I agree to the [Terms of Use](#).

< Back Register

Site licence member
A site licence key is required.

Please select the desired type of licence.
If you are eligible for an institutional discount, please contact inquiries@sketchengine.co.uk.



Introduction to features

Ondřej Matuška



What is it?

- how language is used
- billion-word corpora
- 80 languages
- 400 corpora
- 20 writing systems
- linguists, lexicologists
- lexicographers
- translators
- terminologists
- teachers, students

Terminology

- **word** / word form
went, going, goes, go
- **lemma**
went, going, goes → **go**
- **tag**
go/**VV**/ goes/**VVZ**/ went/**VVD**/ going/**VHG**/
- **PoS**
- lemma + PoS = **lempos**
went, going, goes → **go-v**
- **-lc**, eg. lemma-lc
Blackberry, blackberry, blackberries, Blackberries → **blackberry**

corpus
specific!



tree (*noun*) Alternative PoS: [verb](#) (15,747) [adjective](#) (3)
 English Web 2013 (enTenTen13) freq = [2,756,030](#) (121.25 per million)

modifiers of "tree"	nouns and verbs modified by "tree"	verbs with "tree" as object	verbs with "tree" as subject
1,313,105 0.48	492,977 0.18	598,811 0.22	444,343 0.16
christmas + 68,425 9.94	trunk + 15,004 9.75	plant + 43,894 10.59	line + 7,055 8.68
palm + 16,834 9.08	16,834 9.08	climb + 12,738 8.90	grow + 16,668 8.06
fruit + 11,235 8.66		surround + 8,773 7.94	

christmas tree (*noun*)
 English Web 2013 (enTenTen13) freq = [68,425](#) (3.01 per million)
 (tree-n filtered by christmas-n)

tree: modifiers of "tree"	tree: nouns and verbs modified by "tree"	tree: verbs with "tree" as object	tree: verbs with "tree" as subject
68,425 1.00	2,603 0.04	22,255 0.33	10,000 0.15
artificial + 1,039 6.59	mistletoe 18 7.25	decorate + 3,540 8.45	ornament
rockefeller + 273 6.59	poinsettia 19 7.21	light + 633 6.15	decorate +
fir + 150 5.83	wreath + 131 7.18	adorn + 192 6.04	light +
capitol + 213 5.45	snowman 40 6.92	recycle + 219 5.85	adorn
lighted + 117 5.38	menorah 14 6.59	twinkle 41 5.49	twinkle
artificial 91 5.19	reindeer 30 6.45	trim + 149 5.32	skirt
living + 114 5.17	pre-lit 7 6.42	erect + 104 5.01	farm
pre-lit 71 5.08	santas 11 6.39	festoon 20 4.71	glow
potted 80 4.92	caroler 8 6.33	discard 85 4.68	water
giant + 389 4.82	tinsel 11 6.28	flock 21 4.57	sparkle
miniature 128 4.80	snowflake 21 5.69		

... from tree	55,768 0.02	indigenous + 289 7.42	evergreen + 111 7.27	leave + 372 9.76	forest
"tree" of ...	50,115 0.02	dormant + 232 7.16	pine + 106 7.18	canopy + 175 9.23	hangman
... on "tree"	42,825 0.03	tall + 795 7.12	right + 146 7.04	fruit + 148 8.71	garden

Selecting a corpus

Sketch Engine

words: 2 % / 100,000,000 days: ∞

Home

- + Create corpus
- + WebBootCaT
- + Upload TMX

Parallel corpora

Compare corpora

My jobs

All jobs

Corpora: Recent My own Featured Parallel **All**

Search: cat

Filter by language: all

Language	Name	Words		
Catalan	caTenTen [2014]	4,189,954,719	i	Q
Catalan	CHILDES Catalan Corpus	209,525	i	Q
English	Oxford Children's Corpus 2015 -- Education	1,323,174	i	Q
Persian	TalkBank Persian (deduplicated)	269,753,238	i	Q

Show old versions of corpora

Sketch Engine

British Nation

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

Simple query:

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

Query type simple lemma ph

Lemma:

Phrase:

Word form:

Character:

CQL:

[Tagset summary](#)

Word sketch 

Lemma: 

Part of speech: 

[Advanced options](#)



Subcorpus: **None (whole corpus)** [info](#)

Minimum frequency: **auto**

Minimum score: **0.0**

Maximum number of items in a grammatical relation: **25**

Sort collocations according to: Score Raw frequency

Show lemma coverage:

Show longest-commonest match:

Cluster collocations:

Minimum similarity between cluster items: **0.15**

modifiers of "tree"		
	<u>1,313,105</u>	<u>0.48</u>
christmas +	<u>68,425</u>	9.94
palm +	<u>38,903</u>	9.65
fruit +	<u>27,858</u>	9.05
oak +	<u>24,589</u>	9.04
pine +	<u>21,983</u>	8.93
olive +	<u>15,093</u>	8.32

modifiers of "tree"		
	<u>1,313,105</u>	<u>0.48</u>
christmas +	<u>68,425</u>	9.94
palm +	<u>38,903</u>	9.65

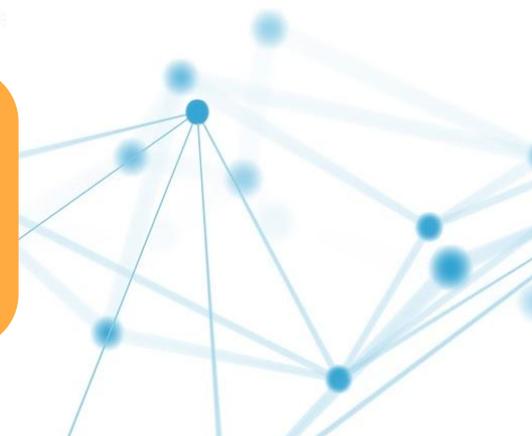
nouns and verbs modified by "tree"		
	<u>492,977</u>	<u>0.18</u>
trunk +	<u>15,004</u>	9.75
branch +	<u>16,834</u>	9.08

verbs with "tree" as object		
	<u>598,811</u>	<u>0.22</u>
plant +	<u>43,894</u>	10.59
climb +	<u>12,738</u>	8.90

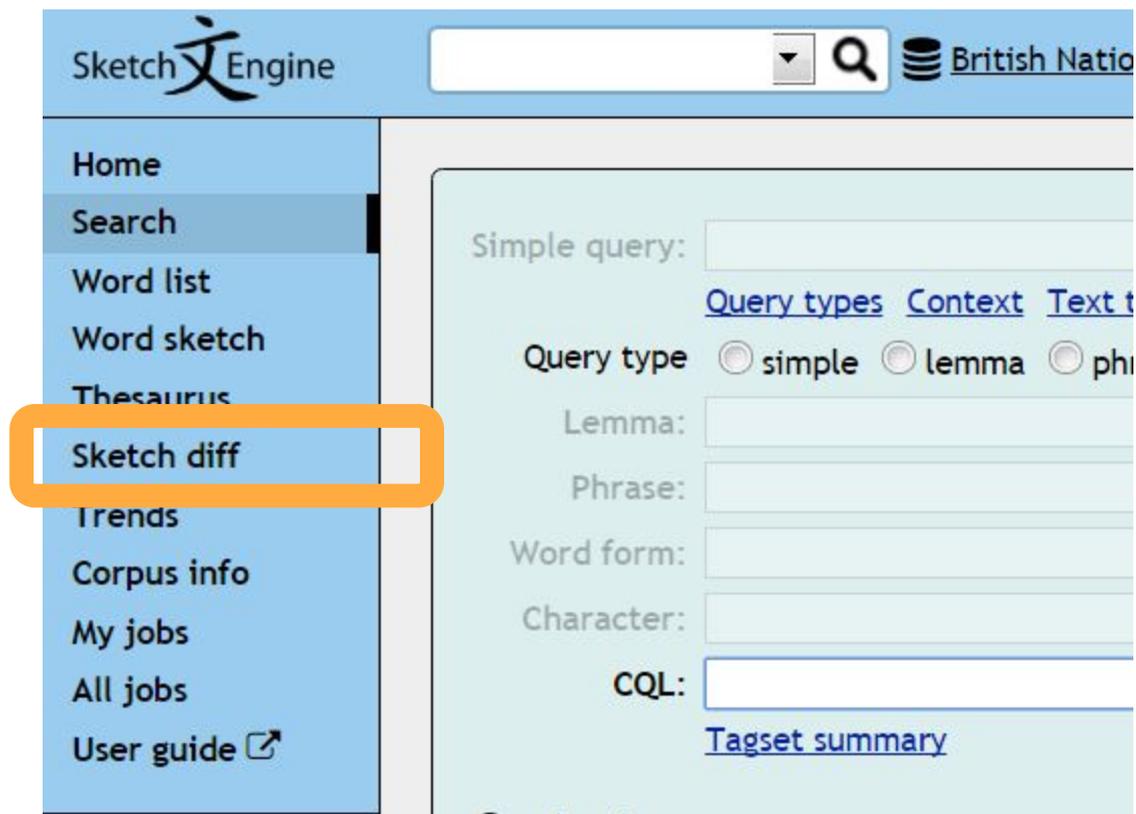
Number of gramrel columns: **5**

Select grammrels:

- All
- ing objects of X
- adjectives after X and noun
- it's X to ...
- objects of Y



Word Sketch Difference



Word Sketch Difference

Word sketch differences ?

Lemma: 

Part of speech: ▼

Sketch diff by: lemma

Second lemma:

subcorpus

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

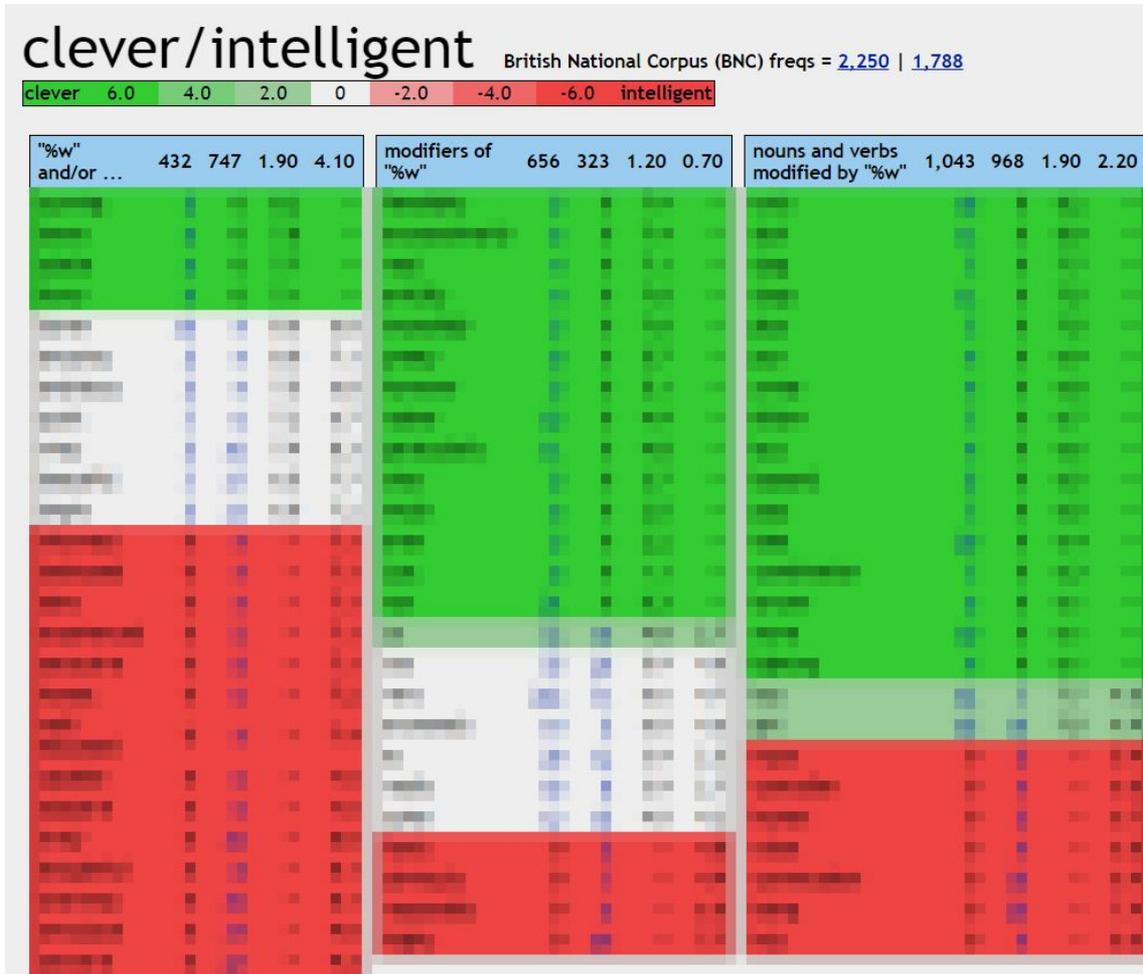
Second subcorpus: ▼ [info](#) [create new](#) ?

word form

First word form:

Second word form:

Word Sketch Difference



"%w" and/or ...	432	747	1.90	4.10	modifiers of "%w"	656	323	1.20	0.70	nouns and verbs modified by "%w"	1,043	968	1.90	2.20
cunning	<u>5</u>	0	8.2	--	fiendishly	<u>5</u>	0	7.9	--	trick	<u>18</u>	0	8.2	--
brave	<u>6</u>	0	7.8	--	extraordinarily	<u>6</u>	0	7.4	--	dick	<u>11</u>	0	8.0	--
subtle	<u>6</u>	0	7.7	--	real	<u>4</u>	0	7.1	--	clog	<u>7</u>	0	7.7	--
funny	<u>4</u>	0	7.1	--	awfully	<u>4</u>	0	7.0	--	chap	<u>11</u>	0	7.5	--
clever	<u>10</u>	<u>4</u>	8.6	6.8	incredibly	<u>5</u>	0	7.0	--	ploy	<u>7</u>	0	7.5	--
amusing	<u>5</u>	<u>4</u>	8.0	7.1	pretty	<u>8</u>	0	5.7	--	pun	<u>4</u>	0	6.9	--
ambitious	<u>5</u>	<u>4</u>	7.5	6.8	however	<u>5</u>	0	5.1	--	swine	<u>4</u>	0	6.8	--
quick	<u>7</u>	<u>6</u>	7.2	6.7	rather	<u>15</u>	0	4.9	--	fellow	<u>7</u>	0	6.7	--
witty	<u>5</u>	<u>10</u>	7.9	8.3	particularly	<u>11</u>	0	4.8	--	lass	<u>4</u>	0	6.6	--
beautiful	<u>7</u>	<u>17</u>	6.5	7.7	that	<u>5</u>	0	3.7	--	bastard	<u>5</u>	0	6.5	--
bright	<u>5</u>	<u>13</u>	6.0	7.2	much	<u>6</u>	0	2.5	--	pass	<u>7</u>	0	6.5	--
informed	0	<u>4</u>	--	7.1	even	<u>5</u>	0	2.2	--	idea	<u>30</u>	0	6.4	--
dedicated	0	<u>4</u>	--	7.1	just	<u>6</u>	0	1.1	--	combination	<u>7</u>	0	6.3	--
alert	0	<u>4</u>	--	7.1	not	<u>31</u>	0	0.3	--	lawyer	<u>6</u>	0	6.1	--
experienced	0	<u>5</u>	--	7.1	so	<u>72</u>	<u>16</u>	4.8	2.7	move	<u>10</u>	0	6.1	--
attractive	0	<u>8</u>	--	7.1	too	<u>75</u>	<u>20</u>	5.9	4.0	lighting	<u>4</u>	0	6.1	--
honest	0	<u>6</u>	--	7.2	very	<u>262</u>	<u>71</u>	6.6	4.7	boy	<u>62</u>	<u>7</u>	7.7	4.6
well-educated	0	<u>4</u>	--	7.4	extremely	<u>14</u>	<u>5</u>	6.0	4.6	girl	<u>65</u>	<u>10</u>	7.5	4.9
capable	0	<u>6</u>	--	7.6	as	<u>30</u>	<u>12</u>	3.9	2.5	input	0	<u>5</u>	--	5.9
adaptive	0	<u>5</u>	--	7.6	really	<u>18</u>	<u>8</u>	3.9	2.7	controller	0	<u>4</u>	--	6.0
lively	0	<u>10</u>	--	8.0	quite	<u>13</u>	<u>15</u>	4.0	4.2	human	0	<u>5</u>	--	6.7
thoughtful	0	<u>8</u>	--	8.1	fairly	0	<u>5</u>	--	4.8	robot	0	<u>4</u>	--	6.8
charming	0	<u>10</u>	--	8.1	obviously	0	<u>7</u>	--	4.9	conversation	0	<u>10</u>	--	6.9
articulate	0	<u>14</u>	--	9.0	reasonably	0	<u>6</u>	--	6.0	being	0	<u>30</u>	--	7.9
sensitive	0	<u>27</u>	--	9.2	highly	0	<u>64</u>	--	7.7	hub	0	<u>9</u>	--	8.0

Word Sketch Difference Advanced options

Advanced options

Separate blocks: all in one block common/exclusive blocks

Minimum frequency:

Maximum number of items in a grammatical relation of the common block:

Maximum number of items in a grammatical relation of the exclusive block:

clever/intelligent British National Corpus (BNC) freqs = 2,250 | 1,788

clever	5.0	4.0	2.0	0	-2.0	-4.0	-6.0	intelligent
"%w' and/or ...	432	747	1.90	4.10				
clever	10	4	8.6	6.8				
many	8	4	4.0	2.9				
ambitious	5	4	8.0	7.1				
quick	7	6	7.2	6.7				
very	5	5	6.4	6.1				
new	4	4	3.1	3.1				
witty	5	10	7.9	8.3				
young	12	18	5.9	6.4				
good	3	8	4.2	4.9				
beautiful	7	12	4.5	7.3				
bright	5	13	6.0	7.2				
modifiers of "%w'	656	323	1.20	0.70				
so	72	16	4.8	2.7				
too	75	20	5.9	4.0				
very	262	71	6.4	4.7				
extremely	14	5	6.0	4.6				
as	30	12	3.9	2.5				
really	18	8	3.9	2.7				
quite	15	15	4.0	4.2				
nouns and verbs modified by "%w'	1,043	968	1.90	2.20				
boy	62	7	7.7	4.6				
girl	65	10	7.5	4.9				
thing	19	5	4.1	2.2				
child	11	6	4.4	3.7				
very	15	11	3.8	3.3				
use	16	15	5.4	5.3				
dog	5	5	5.2	5.2				
man	59	66	5.6	5.8				
software	4	5	4.7	5.1				
woman	19	33	5.1	5.9				
people	26	32	4.2	5.3				
person	6	15	4.3	5.6				
eye	5	14	3.8	5.3				
item	6	19	1.9	3.5				

clever/intelligent British National Corpus (BNC) freqs = 2,250 | 1,788

Common patterns

clever	5.0	4.0	2.0	0	-2.0	-4.0	-6.0	intelligent
"%w' and/or ...	432	747	1.90	4.10				
clever	10	4	8.6	6.8				
many	8	4	4.0	2.9				
ambitious	5	4	8.0	7.1				
quick	7	6	7.2	6.7				
very	5	5	6.4	6.1				
new	4	4	3.1	3.1				
witty	5	10	7.9	8.3				
young	12	18	5.9	6.4				
good	3	8	4.2	4.9				
beautiful	7	12	4.5	7.3				
bright	5	13	6.0	7.2				
modifiers of "%w'	656	323	1.20	0.70				
so	72	16	4.8	2.7				
too	75	20	5.9	4.0				
very	262	71	6.4	4.7				
extremely	14	5	6.0	4.6				
as	30	12	3.9	2.5				
really	18	8	3.9	2.7				
quite	15	15	4.0	4.2				
nouns and verbs modified by "%w'	1,043	968	1.90	2.20				
boy	62	7	7.7	4.6				
girl	65	10	7.5	4.9				
thing	19	5	4.1	2.2				
child	11	6	4.4	3.7				
very	15	11	3.8	3.3				
use	16	15	5.4	5.3				
dog	5	5	5.2	5.2				
man	59	66	5.6	5.8				
software	4	5	4.7	5.1				
woman	19	33	5.1	5.9				
people	26	32	4.2	5.3				
person	6	15	4.3	5.6				
eye	5	14	3.8	5.3				
item	6	19	1.9	3.5				

only patterns

"%w' and/or ...	432	1.90	modifiers of "%w'	656	1.20	nouns and verbs modified by "%w'	1,043	1.90
cunning	5	8.2	fiercely	5	7.9	trick	18	8.2
brave	6	7.8	extraordinarily	6	7.4	dick	11	8.0
subtle	6	7.7	real	4	7.1	clog	7	7.7
funny	4	7.1	awfully	4	7.0	cheap	11	7.5
intelligent	4	6.8	incredibly	5	7.0	play	7	7.5
cool	4	6.7	pretty	6	5.7	pun	4	6.9
dangerous	4	6.4	however	5	5.1	swine	4	6.8
efficient	4	6.4	rather	15	4.9	fellow	7	6.7
bloody	4	6.3	particularly	11	4.8	lass	4	6.6
little	21	6.2	that	5	3.7	bastard	5	6.5
rich	4	5.8	much	6	2.5	pass	7	6.5
strong	6	5.5	even	5	2.2	idea	30	6.4

intelligent only patterns

"%w' and/or ...	747	4.10	modifiers of "%w'	323	0.70	infinitive objects of "%w'	48	1.60
sensitive	22	9.3	highly	68	7.7	understand	4	5.3
articulate	14	9.0	responsibly	6	6.0	know	5	4.2
charming	10	8.1	obviously	7	4.9			
thoughtful	8	8.1	fairly	5	4.8			
lively	10	8.0						
eductive	5	7.6						
capable	6	7.6						
well-educated	4	7.4						
honest	6	7.2						
attractive	6	7.1						
experienced	5	7.1						
alert	4	7.1						
subjects of "%w'	129	9.30						
woman	6	8.3						

Sketch Engine

British Nation

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

Simple query:

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

Query type simple lemma ph

Lemma:

Phrase:

Word form:

Character:

CQL:

[Tagset summary](#)

Bilingual Word Sketch

Lemma:

Part of speech:

[Advanced options](#)

Advanced options

Subcorpus: [info](#) [create new](#)

Minimum frequency:

Number of gramrel columns:

Select gramrels:

<input type="checkbox"/> X * X	<input checked="" type="checkbox"/> X and/or ...	<input type="checkbox"/> X is a ...
<input type="checkbox"/> -ing objects of X	<input type="checkbox"/> ... is a X	<input type="checkbox"/> adjective predicates of X
<input type="checkbox"/> adjectives after X and noun	<input type="checkbox"/> as reflexive	<input type="checkbox"/> in passive
<input type="checkbox"/> it's X to ...	<input type="checkbox"/> modifiers of X	<input checked="" type="checkbox"/> nouns and verbs modified by X

Bilingual word sketch

Language:

Corpus:

Lemma:

Part of speech:

Select gramrels:

<input type="checkbox"/> adj_complement	<input checked="" type="checkbox"/> modifies	<input type="checkbox"/> n_modifier	<input type="checkbox"/> object
<input type="checkbox"/> object_clause	<input type="checkbox"/> object_in	<input type="checkbox"/> object_of	<input type="checkbox"/> predicate
<input type="checkbox"/> subject_np	<input type="checkbox"/> subject_of	<input checked="" type="checkbox"/> y_o	

intelligent (*adjective*) Alternative PoS: [noun](#) (36,903)
 English Web 2013 (enTenTen13) freq = [434,560](#) (19.11 per million)

inteligente (*adjective*)

Use another candidate translation: [vehículo](#) [inteligencia](#) [transporte](#) [integrador](#) [vehículos](#) [carretera](#) [envasar](#) [logística](#) [inclusivo](#)
 Click on collocates to access reciprocal bilingual search

nouns and verbs modified by "intelligent"

	<u>246,094</u>	0.57
being	4,146	7.10
design	13,911	6.82
conversation	2,771	6.80
creature	2,099	6.78
person	5,869	6.12
decision	4,318	6.11
animal	1,752	5.79
discussion	1,664	5.72
robot	607	5.64
discourse	518	5.59
woman	5,005	5.50
designer	1,450	5.47
specie	1,189	5.42
routing	341	5.35
man	6,311	5.34
mind	1,237	5.33
human	596	5.33
lyric	425	5.29
breed	515	5.29
species	479	5.29
individual	1,587	5.18
debate	798	5.14
people	12,465	5.13
agent	1,773	5.12
algorithm	482	5.12

"intelligent" and/or ...

	<u>160,387</u>	0.37
thoughtful	2,222	8.22
articulate	1,451	8.05
witty	1,510	7.94
informed	1,440	7.86
educated	1,146	7.68
capable	1,296	7.52
sensitive	1,442	7.35
rational	1,037	7.26
insightful	967	7.26
funny	1,690	7.23
talented	1,430	7.18
creative	2,375	7.12
smart	1,547	7.07
wise	898	6.90
attractive	1,316	6.88
caring	838	6.79
compassionate	741	6.78
charming	846	6.76
passionate	821	6.74
mature	772	6.72
honest	1,191	6.69
beautiful	3,101	6.66
handsome	627	6.52
curious	559	6.47
motivated	515	6.43

modifies

	<u>180,496</u>	0.54
teléfono	20,671	9.93
tarjeta	4,395	7.39
humor	1,312	6.98
móvil	763	6.27
semáforo	485	6.14
mas	3,650	6.11
hombre	3,932	6.08
edificio	1,634	6.00
sos	459	5.97
ser	1,525	5.93
manera	5,479	5.85
diseño	2,229	5.74
persona	6,693	5.69
robot	339	5.46
sensor	360	5.41
chip	302	5.41
dispositivo	772	5.37
bomba	547	5.30
televisor	277	5.26
etiqueta	340	5.22
máquina	642	5.21
gente	3,200	5.18
contador	278	5.18
medidor	240	5.12
mujer	2,815	5.09

y_o

	<u>24,175</u>	0.07
culto	521	8.24
astuto	233	7.84
sagaz	131	7.28
audaz	194	7.19
conocedor	142	6.86
sabio	213	6.61
valiente	204	6.60
talentoso	133	6.52
estudioso	90	6.51
sensato	128	6.50
perspicaz	72	6.45
ingenioso	103	6.39
sensible	504	6.38
cariñoso	120	6.22
computador	66	6.05
honesto	210	5.99
carismático	79	5.98
simpático	115	5.97
calculador	52	5.91
mordaz	49	5.81
guapo	111	5.77
lúcido	66	5.75
visionario	61	5.74
creativo	436	5.73
pensante	63	5.73

Thesaurus

Thesaurus 

Lemma:

Part of speech: 

[Advanced options](#)



Thesaurus

argue *(verb)*
 English Web 2013 (enTenTen13) freq = 1,269,171 (55.84 per million)

Lemma Score Freq

discuss	0.392	2,763,400
argue	0.362	3,162,502



argue/discuss English Web 2013 (enTenTen13) freqs = 1,269,171 | 2,763,400

argue 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 discuss

"%w" and/or ...	22,661	127,962	0.30	0.50
plead	225	0	7.8	--
yell	321	0	7.6	--
brief	373	10	9.0	1.3
bicker	300	10	8.7	1.3
fight	2,083	45	9.4	2.8
reason	230	28	8.1	2.8
complain	372	48	8.0	3.3
disagree	482	107	8.5	4.6
argue	418	755	8.2	7.4
debate	1,014	5,716	9.5	10.3
explain	172	1,474	5.9	7.9
decide	78	1,177	5.4	7.9
demonstrate	63	1,413	5.3	8.2
vote	41	989	4.8	7.7
agree	136	2,222	5.5	8.5
present	199	4,804	6.0	9.5
examine	27	1,178	3.6	7.8
consider	43	1,621	3.9	8.0
dissect	15	1,469	4.1	8.5
analyze	56	2,135	3.6	8.1
explore	24	2,107	2.4	8.1
read	55	4,704	2.1	8.2
share	36	4,522	2.5	8.8
meet	30	4,072	2.2	8.7
review	11	3,299	1.2	8.7

subjects of "%w"	336,108	350,654	6.60	2.10
petitioner	1,023	0	6.5	--
plaintiff	3,396	19	7.9	0.4
opponent	1,993	18	7.1	0.2
proponent	1,729	21	7.3	0.9
defendant	3,384	41	7.8	1.4
critic	4,978	152	8.3	3.2
supporter	1,410	46	6.6	1.6
prosecutor	1,302	46	6.6	1.7
advocate	1,804	98	7.1	2.8
lawyer	4,085	336	7.6	3.9
attorney	3,136	357	7.3	4.2
economist	1,486	186	6.8	3.8
other	7,375	1,137	6.6	3.9
scholar	1,923	512	7.0	5.1
brief	1,112	327	6.6	4.8
author	5,390	4,601	7.5	7.2
paper	3,320	7,431	7.0	8.1
article	2,416	17,013	5.9	8.7
board	336	2,687	3.5	6.5
participant	277	2,659	3.5	6.8
chapter	214	2,572	3.7	7.3
panelist	83	1,181	2.9	6.7
panel	151	2,718	2.9	7.1
section	89	2,929	2.0	7.0
topic	20	1,866	0.2	6.7

pronominal objects of "%w"	14,969	70,487	1.30	1.90
she	14	0	3.5	--
ours	9	0	2.7	--
myself	87	0	0.7	--
itself	40	0	0.2	--
herself	78	34	2.0	0.7
he	15	10	1.7	0.7
himself	66	74	0.5	0.6
you	1,265	1,581	0.0	0.4
it	12,201	51,530	3.0	5.1
one	94	558	1.0	3.5
ourselves	0	25	--	0.0
me	0	597	--	0.2
themselves	0	97	--	0.6
yourself	0	128	--	0.7
hers	0	9	--	1.4
oneself	0	16	--	1.4
him	0	936	--	1.5
her	0	646	--	2.0
we	0	20	--	2.1
theirs	0	26	--	2.7
them	0	13,860	--	4.5

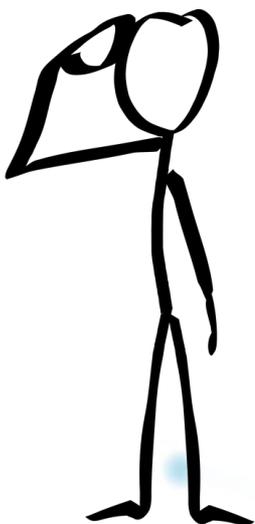
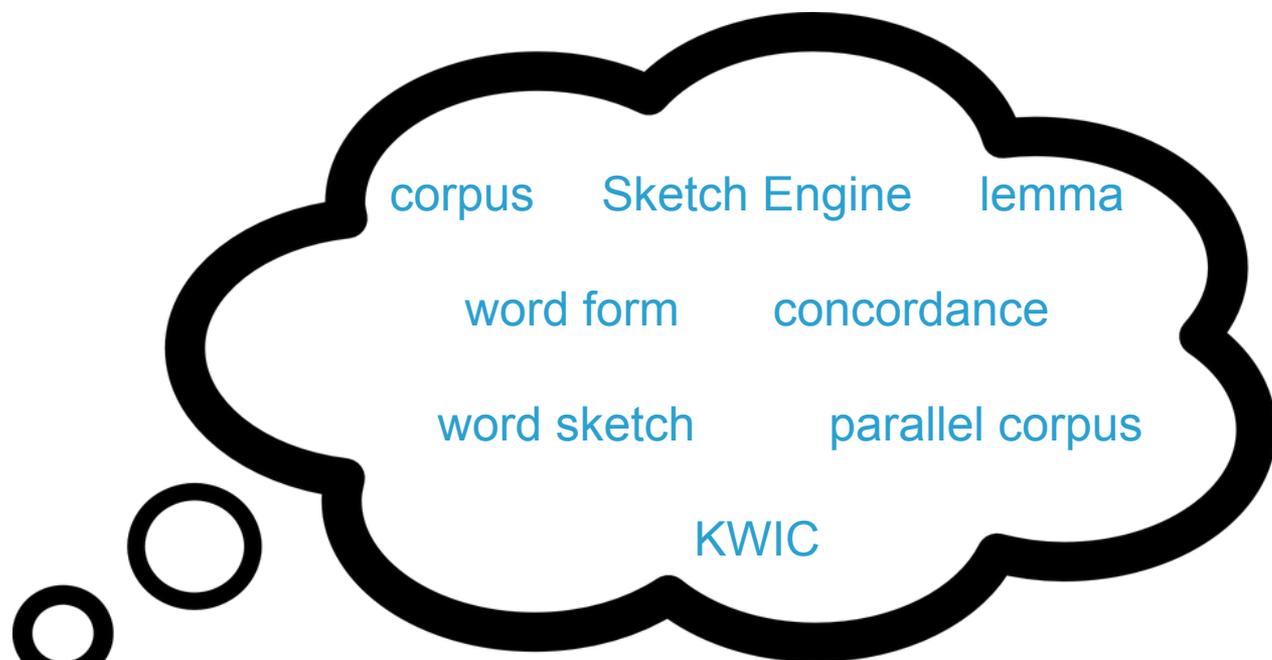
consider 0.300 7,634,698

Create a corpus

- upload files
- gather texts from the internet
- upload TMX (translation memory)

From the internet

- WebBootCaT
- specify the topic
- Sketch Engine does the rest



Corpus from the web

Sketch Engine

words: 2 % / 100,000,000 days: ∞

Home

- + Create corpus
- + WebBootCaT
- + Upload TMX

Parallel corpora

Compare corpora

My jobs

All jobs

Corpora: Recent My own Featured Parallel All

Search: cat Filter by language: all

Language	Name	Words		
Catalan	caTenTen [2014]	4,189,954,719	i	Q
Catalan	CHILDES Catalan Corpus	209,525	i	Q
English	Oxford Children's Corpus 2015 -- Education	1,323,174	i	Q
Persian	TalkBank Persian (deduplicated)	269,753,238	i	Q

Show old versions of corpora

Corpus from the web

WebBootCaT: Create corpus

[Get seed words from Wikipedia](#)

Corpus name

Language 
WebBootCaT is unavailable for languages which cannot be automatically tokenised.

Input type Seed words URLs

Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

Seed words

Random tuples will be selected from the seed words to query a search engine. Input 3 to 20 words or multiword expressions. Use space as separator. Enclose multiword expressions into quotes (").

Compile corpus when finished Automatically compile corpus when WebBootCaT processing is finished.

[Show advanced options](#)



WebBootCaT: Seed words from Wikipedia

Wikipedia URL	Footwear
	https://en.wikipedia.org/wiki/Footwear
Additional Wikipedia URL	stilet
	Stiletto heel
Additional Wikipedia URL	Stiletto
	No Stiletto
Additional Wikipedia URL	Skillet (band)
	Skillet discography
Additional Wikipedia URL	Stille reaction
	Stillerska Filmgymnasiet
Additional Wikipedia URL	Still the Same
	Stiletto (2008 film)
	Still the Same (Slade song)

Cancel < Back Next >

Seed words from Wikipedia

The screenshot shows a web interface for selecting seed words. It features a grid of words with checkboxes and counts. The words are arranged in four columns. The first column has words like 'stiletto (18)', 'Archived (3)', 'Footwear (3)', 'citation (5)', 'Paslawsky (1)', 'Mistinguett (1)', 'pointed-toe (1)', 'slenderness (1)', and 'Heel (2)'. The second column has 'Stiletto (7)', 'Ötzi (2)', 'stiletto (2)', 'semi-stiletto (1)', 'top-piece (1)', 'round-toe (1)', 'Piffle (1)', 'synecdoche (1)', 'Biba (1)', 'stopper', 'opulent (1)', and 'flared (1)'. The third column has 'high-heeled (5)', 'Vivier (2)', 'mass-produced (2)', 'Stiletto-style (1)', 'heavy-feeling (1)', 'wide-shouldered (1)', 'Bergstein (1)', 'Balderdash (1)', 'Stiletto-style (1)', 'adornments (1)', 'adherents (1)', and 'sharpness (1)'. The fourth column has 'heel (22)', 'heels (19)', 'Retrieved (2)', 'fertile-looking (1)', 'cordwainers (1)', 'chopines (1)', 'pattens (1)', 'mass-producing (1)', 'Practitioners (1)', and 'Practitioners (1)'. At the bottom, there are two buttons: '< Back' and 'Use WebBootCaT with selected words'. The interface is overlaid on a background of a network graph with blue nodes and lines.

Corpus name

Language 

WebBootCaT is unavailable for languages which cannot be automatically tokenised.

Input type

Seed words

URLs

Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

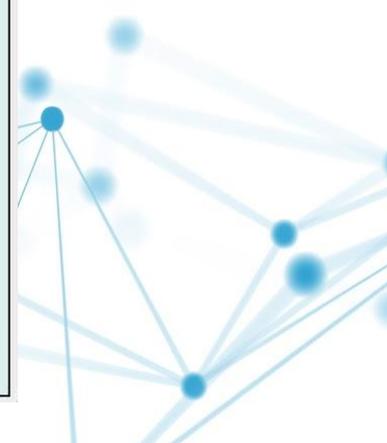
Seed words

Random tuples will be selected from the seed words to query a search engine. Input 3 to 20 words or multiword expressions. Use space as separator. Enclose multiword expressions into quotes (").

Compile corpus when finished

Automatically compile corpus when WebBootCaT processing is finished.

[Show advanced options](#)



Select URLs to download

Query: Footwear pointed-toe stiletto

- [http://www.polyvore.com/black pointed toe stilettos/s?query=black+pointed+toe+stilettos](http://www.polyvore.com/black+pointed+toe+stilettos/s?query=black+pointed+toe+stilettos)
- [http://www.polyvore.com/pointed toe stilettos/shop?query=pointed+toe+stilettos](http://www.polyvore.com/pointed+toe+stilettos/shop?query=pointed+toe+stilettos)
- <http://www.aliexpress.com/popular/pointed-toe-stilettos.html>
- https://www.facebook.com/permalink.php?story_fbid=1727379387499185&id=1577861719117620
- https://www.facebook.com/permalink.php?story_fbid=1727378544165936&id=1577861719117620
- <http://www.target.com/th/pointed+toe+stiletto+shoes>
- <http://www.ebay.com/bhp/pointed-toe-boots>
- <http://www.stilettostyle.com/>
- <http://www.heels.com/shoe-style/pointed-toe>
- <http://www.shopstyle.com/browse?>

footwear: WebBootCaT: Downloading data...



Successfully processed files	0
Files remaining	56
Data downloaded	225 kB
Tokens retrieved	0
Tokens per file (avg)	0
Time elapsed	0:08
Estimated time remaining	4:08
Average file processing time	4.4 s

Errors	2
- unable to retrieve	0
- invalid content-type	0
- file size out of range	0
- cleaned file size out of range	2
- keywords filter applied	0
- unable to convert to text	0
- duplicate	0

 [Cancel processing](#)

```
Processing http://www.polyvore.com/stilettoes_shoes/shop?query=stilettoes+shoes
- Content-type: text/html; charset=UTF-8
- File type: html
- Data read: 34.9kB
- Detected character encoding: utf_8
- Plain text size: 0 characters
- Too small (min size: 1024 characters)
Processing http://www.yebhi.com/online-shopping/women/stilettoes.html
```

Corpus from web

Sketch  Engine

Ready!

sketchengine.co.uk



Keywords & Terms

Sketch Engine

words: 2 % / 100,000,000 da

Home **1**

- + Create corpus
- + WebBootCaT
- + Upload TMX

Parallel corpora

Compare corpora

My jobs

All jobs

Advanced features

Corpora: Recent **2** My own Featured Parallel All

Search: Filter by language: all

Language	Name	Words	
English	My photography corpus	638,077	3
English	NLP	853,678	
English	tea	247,966	
Spanish	fotografía	194,137	

+ [Create new corpus](#) | + [WebBootCaT](#) | + [Upload TMX](#)

Home

- + Create corpus
- + WebBootCaT
- + Upload TMX

Parallel corpora

Compare corpora

My jobs

All jobs

Advanced features

Search corpus ⓘ

Q Concordance

Q Keywords / terms

Q Thesaurus

Q Sketch-Diff

My photography corpus

my_photography_corpus

+ [Add new file](#) | + [Add data from web using WebBootCaT](#) | ⌚ [Compile corpus](#) | 🔍 [Search corpus](#)

#	Original file	Plain text	Vertical	Tokens ⓘ	Owner
	📁 take_5 (81 files)			329,121	Mr. Ondřej Matuška
	📁 smaz4 (9 files)			8,994	Mr. Ondřej Matuška
	📁 smaz3 (81 files)			205,723	Mr. Ondřej Matuška
	📁 my_photography_corpus (88 files)			226,990	Mr. Ondřej Matuška

My photography corpus: Extracted keywords / terms

[Change extraction options](#) Download singlewords: [TBX](#) [CSV](#). Download multiwords: [TBX](#) [CSV](#).

Singlewords and multiwords are ordered by [keyness score](#). The score and corpus frequency (leading to the respective concordance) are displayed in parentheses. **Highlighted** words were used as seeds in a previous WebBootCaT run within this corpus.

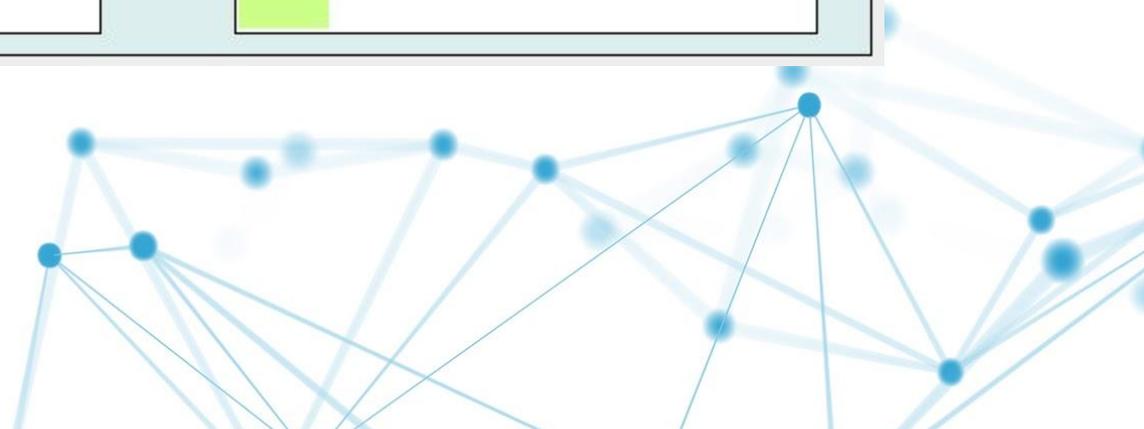
[<< Back to corpus files](#)

Use WebBootCaT with selected words

Extracting singlewords...



Extracting multiwords...



Keywords & Terms

[Change extraction options](#) Download singlewords: [TBX CSV](#). Download multiwords: [TBX CSV](#).

Singlewords and multiwords are ordered by [keyness score](#). The score and corpus frequency (leading to the respective concordance) are displayed in parentheses. **Highlighted** words were used as seeds in a previous WebBootCaT run within this corpus.

[<< Back to corpus files](#)

Use WebBootCaT with selected words

Single-word	Score	F	RefF
<input type="checkbox"/> aperture	W 462.02	1,493	72,603
<input type="checkbox"/> mirrorless	W 443.61	407	4,375
<input type="checkbox"/> photodiode	W 344.07	297	2,790
<input type="checkbox"/> dof	W 337.04	317	5,072
<input type="checkbox"/> nikon	W 313.75	1,037	74,799
<input type="checkbox"/> shutter	W 308.96	1,158	87,859
<input type="checkbox"/> viewfinder	W 294.12	438	21,259
<input type="checkbox"/> iso	W 292.82	2,052	183,977
<input type="checkbox"/> pixel	W 246.40	1,089	107,679
<input type="checkbox"/> lens	W 241.93	3,926	455,860
<input type="checkbox"/> cmos	W 231.10	374	25,088
<input type="checkbox"/> autofocus	W 224.85	283	14,484
<input type="checkbox"/> dslr	W 223.50	492	42,283

Multi-word	Score	F	RefF
<input type="checkbox"/> focal length	W 724.05	626	1
<input type="checkbox"/> shutter speed	W 481.46	416	1
<input type="checkbox"/> image sensor	W 433.00	333	0
<input type="checkbox"/> image quality	W 404.13	426	3
<input type="checkbox"/> image stabilization	W 383.70	295	0
<input type="checkbox"/> optical zoom	W 347.38	267	0
<input type="checkbox"/> default value	W 338.30	260	0
<input type="checkbox"/> dynamic range			
<input type="checkbox"/> manual focus			
<input type="checkbox"/> wide angle			
<input type="checkbox"/> digital zoom			
<input type="checkbox"/> zoom lens			

- Related Wikipedia articles
- [Dynamic range](#)
 - [High-dynamic-range imaging](#)
 - [High-dynamic-range rendering](#)
 - [High dynamic range](#)
 - [Wide dynamic range](#)

Word lists

- all lemmas
- all word forms
- words starting / containing / finishing with ...
- all nouns, verbs
- tags
- n-grams
- search by × display

Subcorpus: [info](#) [create new](#) ?

Search attribute:

use n-grams. Value of n: from to ?

hide/nest sub-n-grams

Filter options:

Filter word list by: Regular expression: ?

Minimum frequency:

Maximum frequency: (0 = no maximum frequency)

Whitelist: No file chosen

Blacklist: No file chosen [format](#)

Include non-words

Output options:

Frequency figures: Hit counts Document counts ARF

Output type: Simple

Keywords

Reference

(sub)corpus

Prefer: rare words common words ?

Change output attribute(s)

You can select one or more output attributes. Please note that this option can be time-consuming.

Words beginning with gn-

Word list

Corpus: English Web 2013 (enTenTen13)

Page [Next >](#)

<u>lemma</u>	<u>Freq</u>
gnome	40,312
gnaw	20,781
gnat	10,634
gnash	9,967
gnarly	9,414
gnocchi	8,634
gnawing	8,440
gnarled	8,028
gnosis	7,600
gnu	6,425
gnostic	4,847
gnc	4,290
gnarl	2,681
gneiss	2,432
gna	1,528
gnomish	1,433
gnosticism	1,336
gn	1,289
gnomon	1,236
gnoll	1,119



Subcorpus: None (whole corpus) [info](#) [create new](#) [?](#)

Search attribute: **lempos (lowercase)**

use n-grams. Value of n: from 2 to 2 [?](#)

hide/nest sub-n-grams

Filter options:

Filter word list by: Regular expression: **.*-n**

Minimum frequency: 5

Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: No file chosen

Blacklist: No file chosen [format](#)

Include non-words

Output options:

Frequency figures: Hit counts Document counts ARF

Output type: Simple

Keywords

Reference: English Web 2013 (enTenTen13)

(sub)corpus: (whole corpus)

Prefer: rare words common words 1 [?](#)

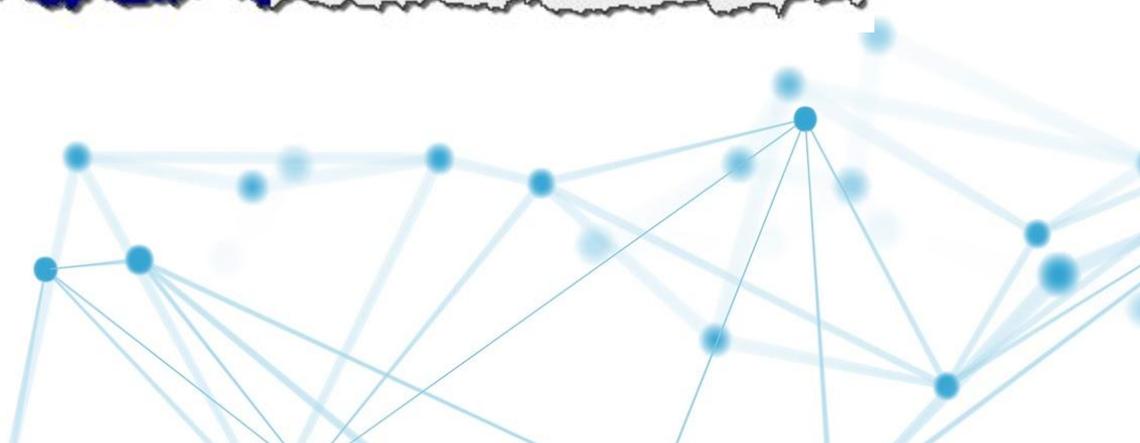
Change output attribute(s)

lemma

You can select one or more output attributes. Please note that this option can be time-consuming.

The most frequent nouns

<u>lemma</u>	<u>Frequency</u>
P N time	63,157
P N year	56,328
P N people	41,811
P N way	39,073
P N man	30,578
P N thing	24,970
P N work	24,785
P N case	23,790
P N part	23,387
P N number	22,556
P N area	22,186
P N course	21,576
P N quest	20,737



n-grams

Subcorpus: [info](#) [create new](#) [?](#)

Search attribute:

use n-grams. Value of n: from to [?](#)

hide/nest sub-n-grams



<u>word (n-grams)</u>	<u>Freq</u>
I do n't know	<u>11,904</u>
the end of the	<u>10,374</u>
at the end of	<u>7,844</u>
I do n't think	<u>6,985</u>
at the same time	<u>4,812</u>
the rest of the	<u>4,712</u>
for the first time	<u>4,712</u>
per cent of the	<u>4,522</u>
as a result of	<u>4,468</u>
at the end of the	<u>3,790</u>
one of the most	<u>3,286</u>
is one of the	<u>3,267</u>
do n't want to	<u>3,266</u>
in the case of	<u>3,245</u>
I do n't want	<u>3,239</u>
to be able to	<u>3,167</u>
the Secretary of State	<u>3,059</u>
On the other hand	<u>2,836</u>
in the form of	<u>2,757</u>
on the basis of	<u>2,743</u>
the top of the	<u>2,673</u>
in the middle of	<u>2,641</u>
do n't know what	<u>2,542</u>
by the end of	<u>2,512</u>
as well as the	<u>2,508</u>
on the other hand	<u>2,460</u>
the way in which	<u>2,426</u>
a member of the	<u>2,415</u>
was one of the	<u>2,329</u>
at the time of	<u>2,288</u>
the middle of the	<u>2,219</u>
a great deal of	<u>2,205</u>

Simple query:

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

Context

Lemma filter

Window: tokens.

Lemma(s): of these items.

PoS filter

Window: tokens.

PoS: adjective adverb conjunction determiner noun noun singular of these items.

- 1 [co.uk](#) loans guaranteed live sort you won thing **rather** than use blame card game. 29. Pay bills
- 2 [hot-brides...](#) `</p><p>` Yes, but I still stand by my post **rather** than the last sentence in her admonition
- 3 [theintelli...](#) sighed and replied, Yes, sometimes. Would you **rather** have you walking the streets in the morning
- 4 [icontempla...](#) will it be placed in your home? Would you **rather** have an upright or cocktail table machine
- 5 [soulfriend...](#) demanded to bring forth maternal instincts **rather** than intelligence, and selflessness rather
- 6 [iefworld.org](#) in helping people live sustainable lives, **rather** than exploiting them for profit. `</p><p>`
- 7 [annenbergc...](#) the speaker chose that particular phrase **rather** than some other phrase? `</p><p>` Marcana and
- 8 [maplestory...](#) lurching into speaking about itself, on **rather** labored terms. Regardless if it developed
- 9 [ownwebsite...](#) to get across to your website's visitor, **rather** than what they will see upon arriving at
- 10 [collisionr...](#) `</p><p>` We need to not limit ourselves, but **rather** we need to open our minds to see the best
- 11 [hotcourses...](#) tutors in office hours whenever possible, **rather** than suffer in silence over a problem you
- 12 [bookmyboot...](#) administrators spent too much time managing servers **rather** than innovation. This technology i.e. the
- 13 [danielamer...](#) real government spending is \$140 billion, **rather** than \$70 billion. To find private sector
- 14 [aircooledt...](#) fan similar to the Porsche 911 kits, but **rather** than the fan being in a vertical configuration
- 15 [ih2012.org](#) made the decision so as to settle the suit **rather** since take it to the Colorado court. `</p>`
- 16 [audioholic...](#) the issue is one of signal transmission **rather** than hardware. `</p><p>` Mitsubishi `</p><p>` The
- 17 [mb-soft.com](#) have obtained her; but otherwise, I had **rather** die than force her against her will. ' `</p>`
- 18 [greencircl...](#) many sorts of little company funding option **rather** than a lot of them are: carry inside a
- 19 [com.au](#) 's my younger brother Daniel, who can be **rather** frustrating and infuriating at times, but
- 20 [org.uk](#) headache, but in a good way! The support helped **rather** than having a negative effect and all three
- 21 [anma.org](#) of life that involve giving and receiving **rather** than getting and taking. Oddly enough,
- 22 [project206...](#) general goals should be sought in parallel **rather** than sequentially. For the most part, learning
- 23 [puliyei.com](#) introduction of the vaccine in the hospitals (**rather** than in outreach sessions) or in selected
- 24 [centauri-d...](#) their paces, though at a separation of 12 **rather** than 40 meters. Quite a lot of serious
- 25 [violeccott...](#) outdoorsy products because I would much **rather** be hanging out in a garden than stuck in
- 26 [paxillawsu...](#) Seroxat, which is not a Paxil generic, but **rather** the European version of the same drug from

Simple query:

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

Query type simple lemma phrase word character CQL

Lemma: PoS:

Phrase:

Word form: PoS: match case

Character:

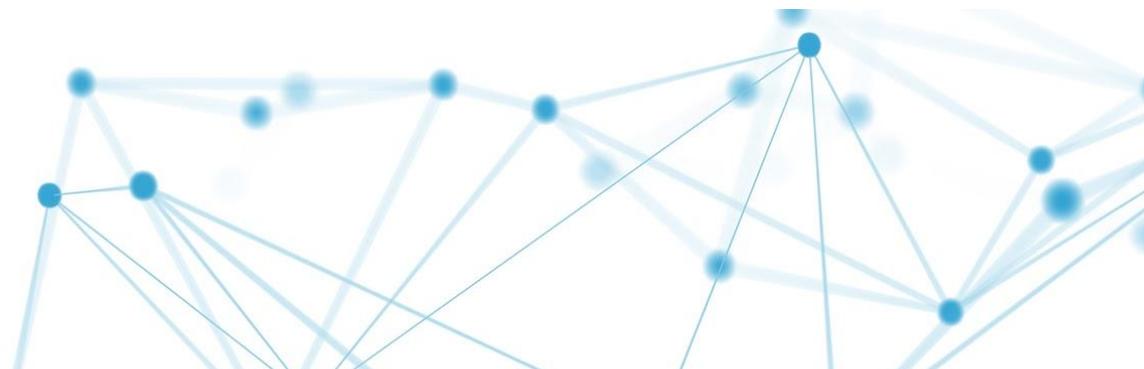
CQL: Default attribute:

[Tagset summary](#)

[lemma!="would"]

[lemma="rather"]

[!(tag="J.*|RB.* | word="than")]



Query **would, rather, RB.?|J.?, than** 8,565 (76.28 per million) 

[First](#) | [Previous](#) Page of 429 [Next](#) | [Last](#)

JOP	pre-eminently German - field of linguistics,	or rather,	to use the contemporary term, philology
JOV	technologies in general or e-mail in particular	; rather,	it is the lack of shared electronic classification
JOV	control over record keeping in government	, rather it	gives advice on selection for archival
JOV	is no one network known as The Internet	; rather,	regional nets like SuraNet, PrepNet, NearNet
JOV	intended to be binding or prescriptive,	but rather to	provide impartial guidance for the perplexed
JOV	nor an unstructured collection of tagsets	. Rather we	offer an extensible framework containing
JOV	source and thus impinge upon interpretation	. Rather,	it provides a set of very flexible solutions
JOV	interested in our data. It will, I suspect,	be rather a	long time before historians adopt SGML-authoring
JOV	about the state of British higher education	. Rather,	it is to indicate that in this particular
JOV	base became much more than the database (or rather its	constituent tables). Between the management
JOT	alternative explanations were readily proffered)	but rather the	failure to find a convincing mechanism
JOT	of magnetization, as earlier suggested,	but rather a	consequence of the direction of magnetization
JOT	between about 175 and 275 km from the trench,	a rather greater	distance than is typically the case with
JOT	direct consequence of continental collision	; rather it	reflects developments that have occurred
JOT	so the literature from this perspective.	is rather limited	. King (1967) is still useful for its descriptions
JOT	0.13 km ³ a ⁻¹ during historic time, and	a rather lower	average of 0.06 km ³ a ⁻¹ over the past
JOT	often branches towards the terminus of a	flow rather like	the distributaries of a river delta. A
JOT	landforms of a similar form but larger size	. Rather volcanoes	can be said to have a morphological capacity
JOT	practice they rarely operate separately	; rather,	the effects of one aid the operation of
JOT	energy within their chemical structures	in rather the	same way that the ball in Figure 6.4 possesses

Features

- Word Sketch
- Word Sketch Difference
- Bilingual Word Sketch Difference
- Thesaurus
- Keywordd & Terms
- Word lists
- n-grams
- Concordance
- CQL



