Collocates at English-Corpora.org / Word Sketches at Sketch Engine

Both the corpora from English-Corpora.org and the corpora from Sketch Engine allow users to quickly and easily find the collocates (nearby words) that occur near a given "node" word, such as *bread*. In the case of Sketch Engine, it divides the collocates grammatically / functionally (such as "modifier of / modified by", or collocate as subject or object of the nearby verb, among other functional categories):

modifiers of "bread"	nouns modified by "bread"	verbs with "bread" as object	verbs with "bread" as subject	"bread" and/or ...	prepositional phrases	adjective predicates of "bread"
unleavened of unleavened bread	crumb bread crumbs	bake baked bread	crumb the bread crumbs	butter bread and butter	... of "bread"	stale bread was stale
rye rye bread	pudding bread pudding	toast toasted bread	bake bread baking	wine the bread and wine	... with "bread"	soggy the bread soggy
crusty crusty bread	dough bread dough	slice thing since sliced bread	slice the bread slices	pasta bread , pasta	... for "bread"	delicious bread is delicious
banana banana bread	pasta bread , pasta	butter bread is buttered	toast bread toasted	cheese bread and cheese	... on "bread"	moist bread moist
sourdough sourdough bread	flour bread flour	leaven leavened bread	cube the bread cubes	cereal breads and cereals	"bread" in ...	unleavened bread was unleavened
garlic garlic bread	roll bread rolls	eat eat bread	dip bread dipped in	pastry breads and pastries	"bread" with ...	fresh bread fresh
wheat whole wheat bread	cereal breads , cereals	consecrate consecrated bread and wine	taste bread tastes	cake bread , cakes	"bread" of ...	tasty bread was tasty
pita pita bread	slice bread slices	dip dipping bread	soak bread soaks	milk bread and milk	"bread" for ...	crusty bread is crusty
					... in "bread"	
					"bread" from ...	
					... as "bread"	
					"bread" on ...	

English-Corpora.org indicates whether the collocates occurs before or after the node word (such as modified by or modifier of), as shown by the highlighted boxes below; for example, *bread and butter* (colored square = *bread* before *butter*) or *loaf of bread* (colored square = *bread* after *loaf*).

COLLOCATES BREAD NOUN See also: VERB Advanced options Collocates Clusters Topics Websites KWIC

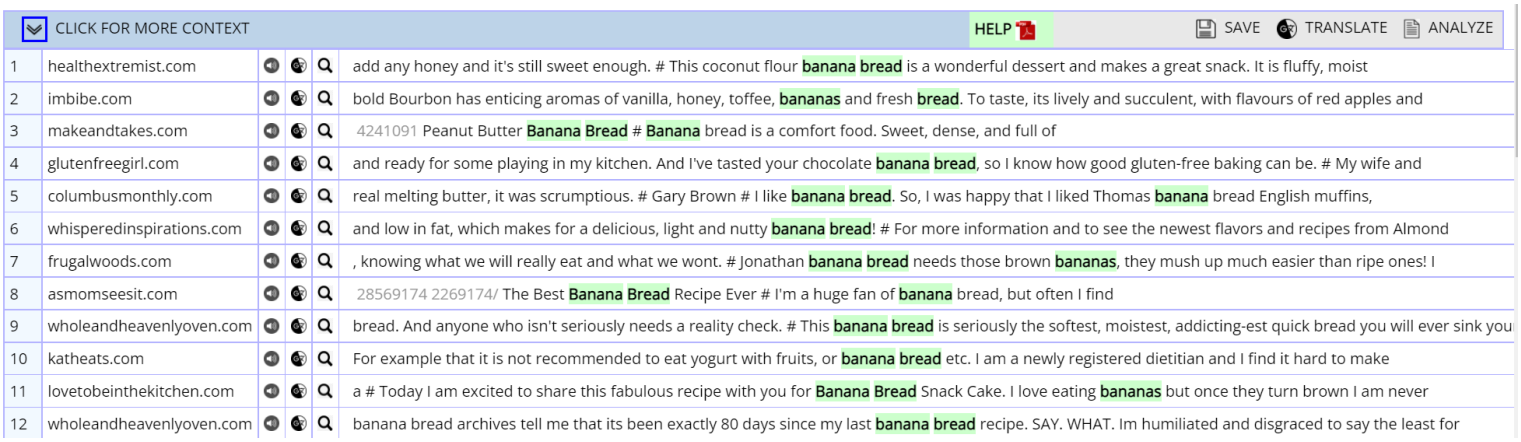
+ NOUN		NEW WORD	?
18101	6.82	butter	
16525	9.53	loaf	
14090	7.59	slice	
11801	7.37	banana	
10912	4.62	recipe	
10595	9.33	crumb	
10267	5.75	cheese	
8447	7.02	wheat	
8324	3.15	piece	
8310	6.47	flour	
7944	4.54	wine	
7421	6.95	pasta	
6171	5.39	grain	

+ ADJ		NEW WORD	?
13416	4.01	white	
9486	4.44	fresh	
8116	3.29	whole	
5963	11.23	unleavened	
5906	7.42	baked	
5848	6.39	homemade	
4879	8.38	sliced	
4751	4.28	french	
4750	9.79	crusty	
3899	3.19	daily	
3793	4.36	delicious	
3690	3.59	sweet	
3401	4.40	brown	

+ VERB		NEW WORD	?
14936	4.18	eat	
14031	6.63	bake	
6225	2.42	serve	
4363	2.46	break	
3300	7.79	toast	
3129	2.02	cut	
2308	3.10	spread	
1854	4.90	dip	
1846	5.35	slice	
1803	2.65	cook	
1698	2.30	rise	
1604	3.65	taste	
1559	2.31	mix	

+ ADV		NEW WORD	?
2509	6.35	freshly	
740	3.62	lightly	
437	3.25	evenly	
285	2.17	traditionally	
137	4.00	thinly	
101	3.85	deliciously	
75	2.15	generously	
74	4.87	thickly	
68	4.38	lengthwise	
45	2.17	alongside	
42	4.01	ever-more	
37	2.60	diagonally	
37	2.65	liberally	

You can easily see the node word / collocates in context, by clicking on the "text" icon (in red above):



CLICK FOR MORE CONTEXT HELP SAVE TRANSLATE ANALYZE

1	healthextremist.com	add any honey and it's still sweet enough. # This coconut flour **banana bread** is a wonderful dessert and makes a great snack. It is fluffy, moist
2	imbibe.com	bold Bourbon has enticing aromas of vanilla, honey, toffee, **bananas** and fresh **bread**. To taste, its lively and succulent, with flavours of red apples and
3	makeandtakes.com	4241091 Peanut Butter **Banana Bread** # **Banana** bread is a comfort food. Sweet, dense, and full of
4	glutenfreegirl.com	and ready for some playing in my kitchen. And I've tasted your chocolate **banana bread**, so I know how good gluten-free baking can be. # My wife and
5	columbusmonthly.com	real melting butter, it was scrumptious. # Gary Brown # I like **banana bread**. So, I was happy that I liked Thomas **banana** bread English muffins,
6	whisperedinspirations.com	and low in fat, which makes for a delicious, light and nutty **banana bread**! # For more information and to see the newest flavors and recipes from Almond
7	frugalwoods.com	, knowing what we will really eat and what we wont. # Jonathan **banana bread** needs those brown **bananas**, they mush up much easier than ripe ones! I
8	asmomseesit.com	28569174 2269174/ The Best **Banana Bread** Recipe Ever # I'm a huge fan of **banana** bread, but often I find
9	wholeandheavenlyoven.com	bread. And anyone who isn't seriously needs a reality check. # This **banana bread** is seriously the softest, moistest, addicting-est quick bread you will ever sink you
10	katheats.com	For example that it is not recommended to eat yogurt with fruits, or **banana bread** etc. I am a newly registered dietitian and I find it hard to make
11	lovetobeinthekitchen.com	a # Today I am excited to share this fabulous recipe with you for **Banana Bread** Snack Cake. I love eating **bananas** but once they turn brown I am never
12	wholeandheavenlyoven.com	banana bread archives tell me that its been exactly 80 days since my last **banana bread** recipe. SAY. WHAT. Im humiliated and disgraced to say the least for

In Sketch Engine, it is not possible to know where the collocate occurs in the "window" of words before or after the node word. For example, with "object of *eat*", both *eat NOUN* and *eat* ***the*** *NOUN* are grouped together. It is possible to manually search for these separately (using the concordance feature), such as *[tag="V.*"] [lemma="bread"]* (*eat bread*) or *[tag="V.*"]* ***[word="the"]*** *[lemma="bread"]* (*eat* ***the*** *bread*). But these searches are very slow and often don't work at all. For example, we searched for [tag="V.*"] [word="the"] [lemma="bread"] (*eat the bread*), and even after 900 seconds (15 minutes) it still didn't show the results.

In the corpora from English-Corpora.org, we can search for either *VERB BREAD* or *VERB the bread* in the 14 billion word iWeb corpus, and it shows the results within about 1-2 seconds (notice the tagging error with *eating disorders*. This happens with Sketch Engine as well, such as *baked / toasted / sliced / buttered bread above*, where it thinks these are verbs).

HELP	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 30,898 | UNIQUE 5,618 +
1	EATING DISORDERS	1351	
2	EAT MEAT	905	
3	EAT LUNCH	687	
4	EAT DINNER	652	
5	EAT BREAKFAST	579	
6	EATING MEAT	466	
7	EATING DINNER	409	
8	EATING LUNCH	367	
9	EAT FOOD	353	
10	EATING BREAKFAST	346	

HELP	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 26,349 | UNIQUE 4,148 +
1	EAT THE BREAD	592	
2	MAKE THE BREAD	565	
3	BAKE THE BREAD	427	
4	TOAST THE BREAD	420	
5	PLACE THE BREAD	382	
6	PUT THE BREAD	370	
7	CUT THE BREAD	342	
8	LET THE BREAD	286	
9	ADD THE BREAD	285	

Or consider the searches for *NOUN BREAD* or *BREAD NOUN* (two word strings), both of which produce results in about 1-2 seconds. In Sketch Engine, searches like [tag="N.*"] [lemma="bread"] will take at least 10-15 minutes (600-900 seconds), if in fact it does ever does finish at all.

HELP	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 83,082 | UNIQUE 4,564 +
1	BANANA BREAD	9524	
2	WHEAT BREAD	4496	
3	GARLIC BREAD	4056	
4	SOURDOUGH BREAD	2824	
5	PITA BREAD	2766	
6	RYE BREAD	2742	
7	CORN BREAD	2223	
8	SODA BREAD	2135	
9	GRAIN BREAD	1997	
10	PUMPKIN BREAD	1927	
11	ZUCCHINI BREAD	1643	

HELP	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 80,315 | UNIQUE 4,302 +
1	BREAD CRUMBS	8772	
2	BREAD RECIPE	4401	
3	BREAD PUDDING	4271	
4	BREAD FLOUR	3431	
5	BREAD DOUGH	2996	
6	BREAD MACHINE	2844	
7	BREAD RECIPES	1747	
8	BREAD KNIFE	1742	
9	BREAD BAKING	1477	
10	BREAD SLICES	1365	

English-Corpora.org is the only site that really allows users to search for complex strings of words, including semantic information. For example, ***POSS =beautiful @BODY*** would search for a possessive (*his, hers, ours*, etc) + a synonym of *beautiful* + any form (=capitalized) of a word in our user-defined "body" list (*face, eyes, hair*, etc). Sketch Engine doesn't allow "semantic" searching like this (synonyms or user-defined word lists). And as we have seen, even when it does allow us to search for strings of words, it is typically hundreds of times as slow as English-Corpora.org.

HELP	ⓘ	★		ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	ⓘ	★	HIS HANDSOME FACE	152	5	12		2	121	8	2	2
2	ⓘ	★	HER BEAUTIFUL FACE	82	5	4	2	4	62	4	1	
3	ⓘ	★	HER LOVELY FACE	48		3	1		40	1	2	1
4	ⓘ	★	HER BEAUTIFUL EYES	44		2	7		35			
5	ⓘ	★	YOUR BEAUTIFUL FACE	44	3	3	25	4	8		1	
6	ⓘ	★	YOUR BEAUTIFUL EYES	33	3	3	17	2	7	1		
7	ⓘ	★	HIS BEAUTIFUL FACE	21	1	3		1	14	1	1	
8	ⓘ	★	HER BEAUTIFUL HAIR	18		1	5	1	10		1	
9	ⓘ	★	HIS BEAUTIFUL EYES	14		3	2	2	5	1	1	
10	ⓘ	★	HIS HANDSOME HEAD	10					9			1
11	ⓘ	★	MY BEAUTIFUL FACE	10			7	2	1			

CLICK FOR MORE CONTEXT HELP SAVE TRANSLATE ANALYZE

1	2019	SPOK	ABC_GMA	talent. MICHAEL-STRAHAN-ABC: (OC His talent. COBIE-SMULDERS-STU: And, like, his handsome face. I was like congratulation on your talent a
2	2019	FIC	PennLitJournal	kidding! " Evelyn said, the reflection of one street lamp beaming down her beautiful face catching her surprised expression. " Natasha's from I
3	2019	MOV	Killer Kate!	that's great. Four girls for Kate. Then I remembered all of your beautiful faces. I don't want to starve you of the action, but I
4	2018	FIC	SouthernRev	exchange for being his friend, and aside from an hour or two of his lovely eyes blinking at them, he'd never give anything back in return. #
5	2018	FIC	NewEnglandRev	Lord's Table; had heard the voice of the Tempter, now saw his beautiful face. Yes, beautiful. Others might say the demon was ugly, repulsive
6	2018	FIC	Fan Fic	Now she grinned, a wicked, devilish grin with sparks of fire in her beautiful eyes. " Let me help you with that! " Before Fred could process
7	2018	NEWS	Los Angeles Times	70% chores, largely unacknowledged, I can see the gloss of youth leave their beautiful eyes. They start to tremble a little, and the words catch
8	2017	FIC	Bk:SirensSong	and Lord Myles in the dining room where they're being detained. # His handsome face barely shows the strain. # Even his skin and clothes, st
9	2017	FIC	Bk:ThroughShadows	woman seemed to deflate at her mother's words, a frown darting across her lovely face. " Not everyone marries these days. It's a new century
10	2017	FIC	Analog	dragon that wound around her body on a field of blue that perfectly matched her lovely eyes. # He sighed. After his involuntary retirement, th
11	2016	FIC	Bk:AmishChristmasGift	'm sorry, Dat. " The girl shook her head, tears filling her striking eyes. " I saw the quilt stand, and I wanted to come see
12	2016	FIC	Bk:AmishChristmasGift	turned back to Naomi. " Frehlicher Grischtdaag. " He smiled, and his handsome face was kind. Yet, there was something sad in his gorgeous e

Returning to the collocates display, at English-Corpora.org it is possible to change the Mutual Information (MI) and/or frequency limits. For example, if we set the MI level low (for example, 1.5) we will get more general words like *break, call,* or *help* as collocates of *enzyme*.

COLLOCATES ENZYME NOUN

SORT BY FREQ MI MIN MI 1.5 GO RESET ? HIDE

+ NOUN		NEW WORD	?
7553	7.40	liver	
7439	4.02	activity	
6088	5.70	protein	
4220	5.42	acid	
4007	2.70	body	
3422	8.26	inhibitor	
3387	3.70	cell	
3251	2.05	level	
2737	2.08	food	
2703	3.38	production	

+ ADJ		NEW WORD	?
12190	9.45	digestive	
2429	9.07	pancreatic	
2409	4.14	involved	
2366	3.58	active	
2195	3.47	responsible	
1950	2.66	natural	
1924	12.06	proteolytic	
1923	2.54	specific	
1611	2.31	certain	
1343	2.33	key	

+ VERB		NEW WORD	?
5818	4.21	break	
5731	3.93	contain	
5345	3.83	produce	
5030	2.38	call	
4533	1.53	help	
3699	8.12	inhibit	
2833	4.74	convert	
2352	7.58	digest	
2237	10.04	catalyze	
2186	5.32	activate	

+ ADV		NEW WORD	?
4868	2.30	down	
731	3.35	naturally	
565	1.56	thus	
445	2.16	normally	
412	1.84	eg	
208	1.66	commonly	
195	1.51	primarily	
185	1.55	mainly	
171	2.72	thereby	
152	1.64	rapidly	

But if we change this to a higher MI score (like 5.0), then the collocates are more specific to *enzyme*:

COLLOCATES ENZYME NOUN

SORT BY FREQ MI MIN MI 5 GO RESET ? HIDE

+ NOUN		NEW WORD
7553	7.40	liver
6088	5.70	protein
4220	5.42	acid
3422	8.26	inhibitor
2691	5.00	vitamin
2448	5.33	restriction
2199	5.46	DNA
1973	5.26	mineral
1897	7.48	digestion
1841	7.44	substrate

+ ADJ		NEW WORD
12190	9.45	digestive
2429	9.07	pancreatic
1924	12.06	proteolytic
1292	7.88	antioxidant
1189	6.48	elevated
947	6.28	metabolic
824	6.21	systemic
715	7.86	hepatic
688	5.67	bacterial
480	7.14	catalytic

+ VERB		NEW WORD
3699	8.12	inhibit
2352	7.58	digest
2237	10.04	catalyze
2186	5.32	activate
1231	8.23	secrete
911	8.49	metabolize
887	6.07	encode
746	6.20	degrade
547	8.90	inactivate
542	6.61	synthesize

+ ADV		NEW WORD
75	5.45	vitro
54	6.87	irreversibly
52	7.47	covalently
43	7.27	superfamily
39	8.24	reversibly
25	5.07	preferentially
23	5.30	synergistically
14	5.38	subfamily
14	6.27	ubiquitously
10	5.71	constitutively

In the previous examples, the collocates were sorted by frequency (with MI just acting as a "filter"). To get even more specific collocates, we can simply sort by Mutual Information itself:

COLLOCATES ENZYME NOUN

SORT BY FREQ MI MIN FREQ 70 GO RESET ? HIDE

+ NOUN		NEW WORD
498	11.57	papain
1004	11.26	lactase
655	11.04	bromelain
679	10.92	amylase
369	10.87	dismutase
374	10.74	catalase
375	10.67	pepsin
994	10.62	reductase
923	10.61	aromatase

+ ADJ		NEW WORD
1924	12.06	proteolytic
218	10.69	rate-limiting
394	10.55	lysosomal
249	10.53	glycolytic
294	10.35	allosteric
160	9.91	cytochrome
12190	9.45	digestive
78	9.15	immobilized
2429	9.07	pancreatic

+ VERB		NEW WORD
2237	10.04	catalyze
172	9.79	protease
215	9.17	denature
117	9.07	metabolise
547	8.90	inactivate
911	8.49	metabolize
208	8.46	pyruvate
1231	8.23	secrete
3699	8.12	inhibit

+ ADV		NEW WORD
75	5.45	vitro
87	4.44	selectively
87	4.40	chemically
102	3.64	genetically
731	3.35	naturally
71	3.31	clinically
171	2.72	thereby
88	2.43	tightly
4868	2.30	down

And at English-Corpora.org, it is possible to have even more control over the collocates. For example, in the following search we find verbs in the past tense that are one or two words to the left of the lemma *enzyme*, where the Mutual Information score is at least 3.5. At Sketch Engine, using the "pre-processed" data, it would not be possible to 1) limit to a specific part of speech (e.g. past tense, or plural nouns), or 2) define the "span" for the collocates, or 3) limit the results to collocates with a particular specificity (e.g. Mutual Information = 3.5 or more). It might be possible to do searches like this with something other than the "pre-processed" results, but it would likely take 600-900 seconds or more – hundreds of times as slow as at English-Corpora.org.

List Word Browse Collocates KWIC

ENZYME Word/phrase [PC

_vvd Collocates [POS]

\+ 4 3 2 1 0 0 1 2 3 4 +

Find collocates Reset

Texts/Virtual Sort/Limit Options

SORTING FREQUENCY

MINIMUM MUT INFO 3.5

HELP			FREQ		ALL	%	MI
1	ⓘ	★	ENCODED	137	46689	0.29	5.67
2	ⓘ	★	CATALYZED	103	5385	1.91	8.37
3	ⓘ	★	PURIFIED	34	30544	0.11	4.27
4	ⓘ	★	SYNTHESIZED	33	20536	0.16	4.80
5	ⓘ	★	INHIBITED	26	17763	0.15	4.67
6	ⓘ	★	IMMOBILIZED	23	5237	0.44	6.25
7	ⓘ	★	IMPLICATED	22	27314	0.08	3.80
8	ⓘ	★	EXCRETED	19	14899	0.13	4.47
9	ⓘ	★	MEDIATED	19	24935	0.08	3.72
10	ⓘ	★	CONJUGATED	13	7022	0.19	5.00
11	ⓘ	★	SECRETED	12	15292	0.08	3.77

In summary, both English-Corpora.org and Sketch Engine can show the collocates from even very large corpora (many billions of words) in just a second or two. But Sketch Engine is mostly limited to this initial, “static”, pre-processed list of collocates. If you want to focus in on specific strings of words (e.g. *NOUN BREAD* or *VERB the BREAD*), or limit the collocates to a particular span before or after the node word, or limit the results to find very specific collocates, then English-Corpora.org will be hundreds of times as fast. And for “semantically-oriented” queries (e.g. *POSS =beautiful @BODY*), English-Corpora.org is essentially the only option.