

Comparing corpora to AI/LLMs I recently released [seven detailed papers](#) (almost 90 pages of information, and more than 100 different tests), which looked at how well the predictions of Large Language Models (or LLMs, like ChatGPT and Gemini) matched up with actual corpus data (see [video summary](#)). Overall, these “white papers” showed that the words and phrases *generated* by LLMs don’t match up very well with actual corpus data. But the LLMs are quite good at *analyzing* data, such as variation between genres, time periods, or dialects.

And where they *really* shine is with [collocates](#) (nearby words) – especially [comparing](#) the meaning and usage of two words via collocates. As I mentioned in the [collocates](#) “white paper”, any corpus site that is based primarily on showing the collocates of words – no matter how advanced these displays may be – might want to re-think their strategy. This is especially true when you consider that LLMs are based on datasets that are a hundred or even a thousand times as large as the largest online corpora. And even more importantly, the neural networks that underly the LLMs are [far more sophisticated](#) than even the most advanced “[association measures](#)” of online corpora, which basically just use different fancy measures of the collocates that are nearby a “node word”.

Integrating corpora and AI/LLMs But perhaps corpora and AI/LLMs should not be “either/or”, but rather “and/with”. In other words, why not take what corpora are really good at – frequency data for words, phrases, collocates, and syntactic constructions (including comparisons between genres, time periods, and genres). And then combine this with the amazing *analytic* and *classifying* capabilities of the LLMs, and their advanced neural network-based knowledge of the relationship between words and phrases. (Of course, while the AI analyses offer powerful insights, users should remember that they represent intelligent suggestions based on patterns in the data, and they are not 100% accurate linguistic conclusions. If you demand absolute perfection, then LLMs might not be for you.)

So this is the approach that we are using at English-Corpora.org. We query the corpora, to get unparalleled frequency data, especially comparisons between genres, time periods, and dialects. In the past, this data would just be presented to the end user, who would need to somehow make sense of the data (and this can be a daunting task for language learners). But now, **with one simple click of a button in the corpus interface, the corpus data is sent (via an API call) to an LLM, where it is analyzed, and then the categorizations and analyses are sent back to English-Corpora.org, where they are then integrated right into the corpus interface.**

For example, the LLM can classify and organize data for collocates and semantically-related phrases, and then display the collocates as part of semantically-relevant “clusters”, each with a label and an explanation. But of course all of the corpora data (including frequency by genre, time period, or dialect) is still there as well. And for the comparisons between words, frequency data in charts, differences between sections of the corpora, or KWIC (Keyword in Context) data, the LLM analyzes the data and sends back a concise explanation of what it thinks the corpus data means. And all of this happens “behind the scenes” and in just a few seconds. Nothing remotely like this is available via any other online corpora.

Range of LLMs Currently, users can select one of six different LLMs – (Chat)GPT (from OpenAI), Gemini (Google), Claude (Anthropic), DeepSeek, Llama (Meta), and Grok (xAI). And as they are using the LLMs to analyze the corpus data (or having the LLMs categorize words, phrases, and collocates into semantically-relevant groups), they can seamlessly move between the different LLMs, and thus get an even wider range of insight.

Timeline for availability The core integration of the LLMs with the corpora at English-Corpora.org is already done. I have presented preliminary information at an invited [Euralex](#) talk in April 2025, and I am giving plenary/keynote talks on this at conferences in [Spain](#) in mid-May 2025 and [Germany](#) in mid-July 2025. The AI/LLM functionality will then be made available to all users of English-Corpora.org soon thereafter; so probably late July 2025. Before making it available to all end users, I still need to refine the framework for “buying credits” for the API calls to the LLMs (English-Corpora.org is charged every time an API call is made), and providing users with access to saved LLM analyses.

Examples In the pages that following, I provide short examples of how the corpus data and LLM analyses and categorizations are integrated together at English-Corpora.org.

1. Classifying collocates (all corpora)

A typical collocates display looks like the following, for **cap** (noun). It’s just a list of words, and it’s up to the end user to decide what the collocates tells us about the meaning and usage of the word.

ON CLICK:

CONTEXT

TRANSLATE (ES)

ENTIRE PAGE

GOOGLE

IMAGE

PRON/VIDEO

BOOK

THESAURUS

 (HELP)

AI:

CATEGORIZE

HELP	ⓘ	★	RE-USE WORDS	FREQ	ALL	%	MI	
1	ⓘ	★	BASEBALL	1447	52303	2.77	7.34	
2	ⓘ	★	SALARY	1360	20366	6.68	8.61	
3	ⓘ	★	TRADE	823	100349	0.82	5.59	
4	ⓘ	★	MARKET	512	184331	0.28	4.02	
5	ⓘ	★	MILLION	463	304182	0.15	3.16	
6	ⓘ	★	ICE	397	78434	0.51	4.89	
7	ⓘ	★	RED	368	169934	0.22	3.67	
8	ⓘ	★	WEARING	349	65746	0.53	4.96	
9	ⓘ	★	SPACE	316	163889	0.19	3.50	
10	ⓘ	★	TAX	291	153160	0.19	3.48	
11	ⓘ	★	BALL	265	90551	0.29	4.10	
12	ⓘ	★	BILLION	236	114903	0.21	3.59	
13	ⓘ	★	BOTTLE	226	22566	0.67	5.20	

But at English-Corpora.org, users can simply click on [**AI: Categorize**] (see above) and then the LLM groups the collocates semantically, and include labels (and explanations) for each of the semantic groups.

ON CLICK:

CONTEXT

TRANSLATE (ES)

ENTIRE PAGE

GOOGLE

IMAGE

PRON/VIDEO

BOOK

THESAURUS

 (HELP)

HELP	ⓘ	★		FREQ	
Clothing & Accessories (This cluster focuses on the physical items worn, or that are otherwise part of the wardrobe of a person. This suggests " relates to the body, or perhaps to a profession.)					
1	ⓘ	★	GOWN n	178	
2	ⓘ	★	CAP n	177	
3	ⓘ	★	STOCKING n	127	
4	ⓘ	★	JACKET n	102	
5	ⓘ	★	GLOVE n	65	
6	ⓘ	★	SHIRT n	62	
7	ⓘ	★	SUNGLASSES n	58	
8	ⓘ	★	T-SHIRT n	50	
Finance & Taxation (This cluster indicates the financial or economic context of "", implying areas like income, costs, and levies, perhaps even personal responsibility.)					
30	ⓘ	★	SALARY n	1383	
31	ⓘ	★	TAX n	336	
32	ⓘ	★	BILLION m	145	
33	ⓘ	★	SPENDING n	134	
34	ⓘ	★	INCOME n	93	
35	ⓘ	★	DEDUCTION n	72	
36	ⓘ	★	PAYROLL n	50	
Extreme Environments (This cluster describes frigid or harsh climates or conditions, such as the arctic environment or things associated with it, which is perhaps reflective of testing extremes.)					
37	ⓘ	★	ICE n	398	
38	ⓘ	★	POLAR j	200	
54	ⓘ	★	ADJUST v	55	
55	ⓘ	★	DOFF v	38	
Tools & Objects (This cluster features everyday objects and tools that might be encountered, implying a broader context, perhaps including the ability to effect change through manual labor.)					
56	ⓘ	★	BOTTLE n	229	
57	ⓘ	★	FEATHER n	135	
58	ⓘ	★	LENS n	101	
59	ⓘ	★	PLASTIC n	72	
60	ⓘ	★	PEN n	54	
61	ⓘ	★	PISTOL n	46	
62	ⓘ	★	CLOTH n	41	

For the COCA and iWeb corpora, you can also group collocates by part of speech (as in this search for *bow*, noun), and you can see whether the collocates typically comes before or after the node word (represented by the gray bars before and after the node word). And at Sketch-Engine, you have even more information about the node/collocate pair. But in no case are the collocates grouped by *meaning*.

But at English-Corpora.org, users can simply click on [**Use AI to group**] and then the LLM groups the collocates semantically, and include labels (and explanations) for each of the semantic groups.

+ NOUN	NEW WORD	+ ADJ	NEW WORD	+ VERB	NEW WORD
Clothing & Adornment (more)		Clothing & Adornment (more)		Precision & Alignment (more)	
862	7.85	tie		308	5.68
104	6.65	ribbon		46	3.81
53	3.42	shirt		44	6.40
44	3.11	dress		23	4.43
41	7.09	satin		17	6.22
20	5.02	lace		16	2.23
18	2.52	belt		11	3.29
17	2.37	jacket		10	3.54
16	4.76	blouse		9	2.52
12	3.24	collar		Weapons & Archery (more)	
11	2.73	skirt		213	3.77
Naval & Maritime (more)		Naval & Maritime (more)		Emotional States & Expression (more)	
175	4.69	boat		19	2.90
140	4.05	ship		18	2.02
124	4.30	wave		17	2.15
116	7.20	stern		12	2.23
114	5.41	port		10	2.02
91	8.93	starboard		9	2.69
41	8.20	thruster		8	2.33
35	3.78	deck		Hunting & Warfare (more)	
30	4.15	anchor		32	3.98
27	3.47	vessel		11	2.36
22	2.09	crew		Musical & Artistic (more)	
21	5.19	hull		11	2.77
20	3.31	rail		8	2.23
11	2.44	cabin		14	4.88
9	4.75	tug		Supplementary Elements (more)	
9	3.81	ferry		275	2.68
Weapons & Archery (more)		Emotional States & Expression (more)		Supplementary Elements (more)	
961	9.36	arrow		16	3.81
201	4.23	shot		11	4.84
62	4.90	sword		Supplementary Elements (more)	
61	2.11	gun		275	2.68

3. Classifying phrases

LLMs can also group phrases semantically, such as *soft NOUN* in COCA (note the genres, and note also that some entries have been removed for reasons of space in this document.)

HELP	①	★		ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
Physical Texture & Sensation (This cluster highlights the tactile and sensory aspects of "soft," emphasizing things that feel gentle and pleasant to the touch or that lack harshness in their texture.)																		
1	①	★	soft tissue	2111	62	67	74	39	36	100	35	707	138	230	146	163	141	173
2	①	★	soft skin	381	9	24	28	6	114	19	4	3	20	30	35	37	24	28
3	①	★	soft touch	378	22	26	28	15	43	51	25	3	32	30	39	24	17	23
4	①	★	soft hands	313	10	11	36	6	60	24	18	2	21	21	19	27	30	28
5	①	★	soft leather	309	2	5	6		109	28	8		28	19	22	33	29	20
6	①	★	soft rock	177	2	21	19	4	18	12	21	3	11	10	18	19	13	6
7	①	★	soft hair	157	2	9	8	2	52	6	3	2	15	11	13	15	12	7
Sound & Voice Quality (This grouping showcases the auditory dimension of "soft," describing sounds and voices that are gentle, subdued, and non-abrasive, creating a sense of calmness or intimacy.)																		
20	①	★	soft voice	1041	12	39	10	11	351	53	52	18	86	86	90	94	87	52
21	①	★	soft music	318	7	27	39	5	48	27	15	8	38	22	24	23	16	19
22	①	★	soft sound	166	2	8			71	5	1	1	19	8	18	13	11	9
23	①	★	soft laugh	110	1	5			46	5	1		7	6	9	6	11	13
24	①	★	soft voices	107	1	2	2	2	38	4	1	5	3	11	11	9	10	8
25	①	★	soft sounds	98	3	3	1	4	25	11	3	2	7	11	7	10	6	5
Visual Appearance & Light (This cluster focuses on the visual qualities associated with "soft," particularly in terms of lighting, colors, and focus, creating a gentle and diffused aesthetic.)																		
27	①	★	soft light	447	22	25	9	1	121	38	21	10	42	44	34	33	32	15
28	①	★	soft glow	266	3	17			79	31	12	1	14	18	15	24	27	25
29	①	★	soft pastels	137	1				2	57	7	2	23	13	18	12		2
30	①	★	soft focus	128	6	10	4	6	16	10	6	14	15	7	16	4	5	9
31	①	★	soft lighting	92	2	2	1	1	9	18	14	1	4	11	6	12	8	3
32	①	★	soft colors	88	4	2		4	11	17	5	4	8	10	10	6	4	3

Or *NOUN business*:

HELP	①	★		ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
Entertainment Business (This cluster highlights the commercial aspect of various entertainment sectors, emphasizing their function as profit-generating industries rather than purely artistic endeavors. It underscores the business side of creativity.)																		
1	①	★	show business	3434	77	71	365	610	96	273	271	28	373	355	345	209	182	179
2	①	★	music business	1910	163	199	110	178	40	235	186	25	116	132	159	153	124	90
3	①	★	movie business	932	30	42	82	117	42	78	105	6	72	88	92	64	58	56
4	①	★	entertainment business	640	40	42	25	80	8	79	79	8	51	86	43	42	32	25
5	①	★	film business	316	24	22	21	35	7	34	34	4	24	26	27	22	16	20
Financial & Real Estate Business (This cluster showcases the financial sector and real estate market, pinpointing "business" as transactions of money and goods for a profit. It describes activities involving investment, lending, and management of assets.)																		
9	①	★	insurance business	509	38	27	17	42	21	52	79	11	65	40	35	34	34	14
10	①	★	estate business	505	27	38	23	52	19	21	91	14	66	35	24	31	24	40
11	①	★	banking business	385	27	24	12	36	6	33	74	6	45	25	51	26	8	12
12	①	★	mortgage business	180	13	11	2	6	1	26	43		6	6	14	36	7	9
13	①	★	securities business	164	5	3	1	4		17	43	13	23	34	9	6	6	
14	①	★	brokerage business	151	4	7		8	1	24	36	1	20	19	12	10	3	6
Media & Communication Business (These phrases reflect the industries involved in disseminating information and creating content for public consumption, defining business as the trade and operations of communications. It focuses on the business of creating, packaging, and delivering information.)																		
17	①	★	news business	784	26	28	16	225	11	40	70	3	49	62	56	69	70	59
18	①	★	newspaper business	394	28	48	15	32	16	23	69	4	39	31	20	25	27	17
19	①	★	publishing business	283	33	18	10	13	16	30	37	10	26	30	17	15	20	8
20	①	★	advertising business	261	22	17	9	20	5	39	32	6	23	12	33	12	10	21
21	①	★	media business	257	34	31	1	25	1	29	39	1	20	12	12	13	16	23
22	①	★	ad business	111	15	8	2	2	3	22	15		11	2	5	7	1	18
Retail & Consumer Business (The phrases reveal the businesses involved in selling products and services directly to consumers, highlighting business as the direct point of sale. It covers both physical goods and services catering to individual needs and preferences.)																		
23	①	★	restaurant business	794	44	26	42	83	30	44	155	8	59	84	58	58	61	42
24	①	★	car business	386	18	12	12	53	9	48	52	4	36	46	20	20	29	27
25	①	★	food business	369	22	11	13	30	8	42	61	14	33	36	16	29	27	27
26	①	★	home business	358	122	32	3	3		63	29	4	22	41	14	6	17	2
27	①	★	catering business	304	17	7	21	20	19	20	58	2	13	25	39	30	18	15

4. Comparing the collocates of two words (to see differences in meaning and usage)

With both English-Corpora.org and Sketch Engine, users can see the contrasting collocates of two words, which provides insight into differences in the meaning and usage of the two words, such as nouns with quandary and predicament – two words with very similar meanings. But again, it’s still up to the end user to figure out what the two contrasting lists mean.

SEE CONTEXT: CLICK ON NUMBERS (WORD 1 OR 2) [\[HELP...\]](#)

Use AI to compare

SORTED BY RATIO: CHANGE TO [FREQUENCY](#)

WORD 1 (W1): **QUANDARY** (6,236) (0.30)

WORD 2 (W2): **PREDICAMENT** (20,690) (3.32)

	WORD	W1	W2	W1/W2	SCORE		WORD	W2	W1	W2/W1	SCORE
1	VERITABLE	20	0	40.0	132.7	1	PRECARIOUS	53	0	106.0	31.9
2	ETHICAL	101	8	12.6	41.9	2	EXACT	71	1	71.0	21.4
3	MORAL	231	31	7.5	24.7	3	DIRE	127	2	63.5	19.1
4	PHILOSOPHICAL	36	5	7.2	23.9	4	UNFORTUNATE	58	1	58.0	17.5
5	DIRECT	15	5	3.0	10.0	5	STRANGE	55	1	55.0	16.6
6	ETERNAL	15	6	2.5	8.3	6	TRAGIC	26	0	52.0	15.7
7	CLASSIC	12	5	2.4	8.0	7	PERILOUS	25	0	50.0	15.1
8	LOGICAL	14	6	2.3	7.7	8	DESPERATE	49	1	49.0	14.8
9	ESSENTIAL	12	9	1.3	4.4	9	FISCAL	24	0	48.0	14.5
10	BIGGEST	14	13	1.1	3.6	10	BAD	46	1	46.0	13.9
11	BIG	14	13	1.1	3.6	11	CURRENT	1104	26	42.5	12.8
12	REAL	84	99	0.8	2.8	12	AWFUL	42	1	42.0	12.7
13	OLD	27	33	0.8	2.7	13	AWKWARD	41	1	41.0	12.4
14	ENVIRONMENTAL	16	22	0.7	2.4	14	UNAWARE	20	0	40.0	12.1
15	EMOTIONAL	13	18	0.7	2.4	15	EGO-CENTRIC	18	0	36.0	10.9
16	EXISTENTIAL	30	42	0.7	2.4	16	EMBARRASSING	35	1	35.0	10.5

Only at English-Corpora.org can users have AI help explain the difference between the two words:

GOOGLE / GEMINI

★★★★★

Collocates of *quandary* The collocates of "quandary" suggest that it frequently involves intellectual, moral, or abstract dilemmas. Words like "ethical," "moral," "philosophical," "logical," and "metaphysical" emphasize this aspect. Adjectives like "unresolved," "eternal," "perpetual," and "age-old" suggest the enduring and complex nature of these dilemmas. The presence of "academic" further reinforces the intellectual connotation. While "biggest," "growing," and "deep" indicate the significance of the problem, the overall impression is that "quandary" is associated with intellectual or ethical decision-making rather than immediate, practical danger. The presence of "beautiful" and "comical" suggest that the word can be used in less serious, more reflective ways.

Collocates of *predicament* The collocates of "predicament" paint a picture of a challenging and often negative situation. Words like "dire," "tragic," "perilous," "desperate," "awful," "grim," "unpleasant," and "hopeless" clearly indicate the unfavorable nature of the circumstances. Adjectives like "awkward" and "embarrassing" point to social discomfort. Several collocates, such as "fiscal," "ecological," "international," and "Palestinian," suggest that predicaments can be related to specific real-world issues. The presence of words like "helpless," "aware," and "conscious" indicates that the person in the predicament is often aware of their difficult situation, potentially lacking the power to easily escape it.

Comparison of collocates of *quandary* and *predicament* The collocates clearly delineate the distinction between "quandary" and "predicament." "Quandary" is strongly associated with intellectual and moral dilemmas, as evidenced by collocates like "ethical," "moral," and "philosophical." The presence of words like "eternal" and "age-old" suggests that these quandaries are often long-standing and complex. In contrast, "predicament" describes a difficult, often negative or dangerous situation. The collocates, such as "dire," "tragic," and "perilous," emphasize the unfavorable nature of the circumstances. Furthermore, collocates like "fiscal," "ecological," and "international" indicate that "predicament" is often used to describe real-world problems, while "quandary" tends to be more abstract and theoretical. The collocates of "quandary" also suggest a more formal register than those of "predicament."

OPENAI / GPT

★★★★★

Collocates of *quandary* The collocates associated with 'quandary' reveal its nuanced and often intellectual nature. Words like 'ethical', 'philosophical', and 'moral' indicate that quandaries often involve complex decision-making scenarios that engage deeper values and principles. Additionally, terms such as 'infinite', 'eternal', and 'metaphysical' suggest a timeless or abstract quality, emphasizing the profound nature of the dilemmas faced. The presence of adjectives like 'creative' and 'fun' also hints at the possibility of positive outcomes or learning experiences, suggesting that a quandary can stimulate critical thinking and exploration, rather than solely presenting a challenge.

Collocates of *predicament* The collocates of 'predicament' illustrate a more urgent and negative connotation compared to 'quandary'. Words like 'dire', 'tragic', and 'desperate' convey a sense of severe difficulty or distress, highlighting the precariousness of the situation. Many collocates, such as 'unfortunate', 'awkward', and 'embarrassing', emphasize the unpleasantness and social implications of being in a predicament. The frequent use of terms like 'fiscal' and 'ecological' suggests that these situations can often be tied to real-world issues, further emphasizing the practical and sometimes urgent nature of a predicament, which contrasts sharply with the more abstract and reflective quality of a quandary.

Comparison of collocates of *quandary* and *predicament* The collocates of 'quandary' and 'predicament' illustrate their differing meanings and usages effectively. 'Quandary' is associated with intellectual, ethical, and philosophical dimensions, inviting reflection and critical thinking. The presence of terms that suggest timelessness or creativity indicates that quandaries can also lead to personal growth or deeper understanding. In contrast, 'predicament' is linked with urgency and severity, often reflecting immediate and challenging situations that require quick resolution. The collocates indicate a more informal and relatable tone in 'predicament', whereas 'quandary' maintains a formal and abstract quality. This distinction underscores that while both terms refer to difficulties, 'quandary' leans towards moral contemplation, whereas 'predicament' emphasizes tangible challenges and distress.

5. Analyzing differences between *specific* sections (genres, time periods, or dialects)

At English-Corpora.org, you can select specific genres, time periods, or dialects, and see which words, phrases, or collocates are more common in one section than another. For example, the following are nouns that are more common with *chain* (noun) in fiction (left) and in academic (right) in COCA. But again, without AI/LLMs, it is up to the corpus user to be a “junior linguist” and figure out what the two sets of words, phrases, or collocates mean.

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION) SEE # TEXTS [\[HELP...\]](#)

Use AI to compare

SEC 1 (FICTION): 118,322,084 WORDS							SEC 2 (ACADEMIC): 119,790,456 WORDS						
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	DOOR	77	2	0.7	0.0	39.0	1	COMMODITY	142	0	1.2	0.0	118.5
2	NECK	307	8	2.6	0.1	38.9	2	MANAGEMENT	102	1	0.9	0.0	100.7
3	LEATHER	26	1	0.2	0.0	26.3	3	VALUE	120	0	1.0	0.0	100.2
4	HEAD	30	0	0.3	0.0	25.4	4	ANALYSIS	74	1	0.6	0.0	73.1
5	GOLD	402	18	3.4	0.2	22.6	5	SUPPLY	562	10	4.7	0.1	55.5
6	FENCE	111	5	0.9	0.0	22.5	6	ORIENTATION	55	0	0.5	0.0	45.9
7	FINGERS	26	0	0.2	0.0	22.0	7	PRODUCTION	47	0	0.4	0.0	39.2
8	LEG	21	1	0.2	0.0	21.3	8	METHOD	44	0	0.4	0.0	36.7
9	CEILING	25	0	0.2	0.0	21.1	9	COVENANT	36	0	0.3	0.0	30.1
10	MAN	40	2	0.3	0.0	20.2	10	MIGRATION	32	0	0.3	0.0	26.7
11	WATCH	72	4	0.6	0.0	18.2	11	REALIST	30	0	0.3	0.0	25.0
12	CROSS	36	2	0.3	0.0	18.2	12	REPEAT	29	0	0.2	0.0	24.2
13	NECKS	18	1	0.2	0.0	18.2	13	TRANSPORT	24	1	0.2	0.0	23.7
14	PADLOCK	21	0	0.2	0.0	17.7	14	PRODUCT	27	0	0.2	0.0	22.5
15	POCKET	20	0	0.2	0.0	16.9	15	MYOSIN	27	0	0.2	0.0	22.5
16	ANKLE	20	0	0.2	0.0	16.9	16	MODEL	26	0	0.2	0.0	21.7

By clicking on [Use AI to compare], the user can have the LLM explain the difference, such as in the following.

OPENAI / GPT

★★★★★

Data from FICTION The collocates of 'CHAIN_n' in fiction reveal a rich tapestry of imagery and themes associated with physical and symbolic connections. Words like "door," "neck," "leather," and "gold" suggest tangible objects that evoke feelings of bondage, security, or luxury. The presence of items like "padlock," "keys," and "fence" hints at themes of restriction and access, while "man," "watch," and "medallion" introduce human elements and personal significance. Overall, the collocates in fiction emphasize the emotional and material associations of chains, often linking them to concepts of identity, power, and entrapment.

Data from ACADEMIC In academic contexts, the collocates of 'CHAIN_n' reflect a more abstract and systematic understanding of the term. Words such as "commodity," "management," and "value" indicate a focus on economic and organizational frameworks. The presence of terms like "analysis," "supply," and "production" suggests a methodical approach to studying chains in terms of efficiency and logistics. Additionally, scientific terms like "myosin," "polymerase," and "dynein" point to chains in biological and chemical contexts, highlighting their functional roles. Thus, in academia, 'CHAIN_n' is associated with structured, analytical discourse rather than the emotional or narrative qualities found in fiction.

Comparison of phrases in [FICTION] and [ACADEMIC] The collocates of 'CHAIN_n' in fiction predominantly emphasize physicality and emotional resonance, focusing on personal and material connections such as "neck," "fingers," and "padlock." In contrast, academic collocates prioritize analytical and systemic aspects, with terms like "management," "analysis," and "production" suggesting a framework for understanding chains in practical or theoretical contexts. This distinction highlights the difference between narrative and analytical uses of the term, where fiction captures human experiences and emotions, while academia seeks to dissect and understand the mechanics and implications of chains in various fields.

ANTHROPIC / CLAUDE

★★★★★

Data from FICTION In FICTION, the collocates of 'CHAIN_n' predominantly relate to physical, tangible objects and bodily associations. Many collocates refer to jewelry or personal adornments (gold, medallion, pendant, watch), restraints or security (padlock, fence), and body parts (neck, fingers, ankle). The collocates suggest that chains in fiction are often used to describe concrete objects that can be worn, secure things, or restrict movement, reflecting their role in creating vivid, physical descriptions in narrative contexts.

Data from ACADEMIC In ACADEMIC texts, 'CHAIN_n' collocates reflect abstract concepts and scientific processes. The terms relate to business processes (commodity, management, supply), scientific concepts (polymerase, dynein, carbon), and analytical frameworks (analysis, model, system). These collocates indicate that chains in academic writing are often used metaphorically to describe sequences, relationships, or systematic connections in various fields like business, chemistry, and biology.

Comparison of phrases in [FICTION] and [ACADEMIC] The key difference between FICTION and ACADEMIC collocates lies in their level of abstraction and domain specificity. Fiction collocates are concrete, physical, and often related to personal objects or bodily restraints, serving narrative and descriptive purposes. In contrast, academic collocates are predominantly abstract and technical, focusing on processes, systems, and scientific concepts. While fiction uses 'chain' to create tangible imagery and emotional resonance, academic usage employs it as a technical term or conceptual metaphor for describing complex relationships and processes in specialized fields.

6. Analyzing differences across *all* sections (genres, time periods, or dialects)

English-Corpora.org shows the frequency of individual words, phrases, collocates, or syntactic constructions in the different sections of the corpora, including genres (e.g. COCA or BNC), time periods (e.g. COHA, NOW, TIME, or EEBO), or dialects (e.g. GloWbE and NOW). But again, without AI it is up to the end user to make sense of this data, such as for *ADJ food* during the last 200 years in COHA:

ON CLICK:

CONTEXT

TRANSLATE (ES)

ENTIRE PAGE

GOOGLE

IMAGE

PRON/VIDEO

BOOK

THESAURUS (HELP)

AI:

CATEGORIZE

ANALYZE

HELP	①	★		ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
1	①	★	GOOD FOOD	781	1	2	1	8	17	15	26	21	24	37	46	57	74	76	55	68	58	74	68	53
2	①	★	OTHER FOOD	283	5	6	12	4	15	15	15	8	18	20	18	13	24	15	16	22	11	15	13	18
3	①	★	FAST FOOD	281																13	29	59	95	85
4	①	★	CHINESE FOOD	259									1		2	1	5	14	15	41	47	39	39	55
5	①	★	LITTLE FOOD	174	2	3	4	4	1	7	2	10	7	9	16	17	18	12	11	14	13	6	11	7
6	①	★	DAILY FOOD	155	5	6	10	14	14	13	16	15	14	9	11	7	3	1	6	4	1		5	1
7	①	★	SOLID FOOD	152	1	2	3	6	10	8	4	7	10	3	11	5	10	11	13	11	8	8	13	8
8	①	★	BETTER FOOD	146		3	3	2	4	6	4	10	7	12	7	8	15	15	8	6	8	4	9	15
9	①	★	HOT FOOD	143			1		2	1		2	2	5	6	14	7	12	15	16	16	15	20	9
10	①	★	REAL FOOD	139							1		3	5	2	5	2	6	5	9	30	17	19	35
11	①	★	AMERICAN FOOD	135							1		2	5	4	4	18	22	11	28	14	6	9	11
12	①	★	BAD FOOD	133		3	4	2	6	7	5	4	5	7	9	10	11	12	6	11	5	10	6	10
13	①	★	BEST FOOD	131	1	2	1	1	3	3	4	4	6	11	3	5	20	6	7	4	9	9	18	14
14	①	★	PROPER FOOD	131	4	6	9	6	8	7	8	16	4	8	13	7	12	5	2	4		6	3	3
15	①	★	PURE FOOD	127						1	1	2	28	18	40	15	3	7	4	3	2	1	2	
16	①	★	FAVORITE FOOD	126	1	1		6	1	3	1	1	3	3	6		2	7	3	5	13	17	28	25
17	①	★	NEW FOOD	124	2	1	1	1	2		4	1	9	5	5	6	4	9	9	8	9	9	16	23
18	①	★	ONLY FOOD	124	1	8	5	6	7	3	3	8	6	5	10	8	10	7	6	8	6	8	5	4
19	①	★	HUMAN FOOD	115		1	4		13	1	5	2	5	7	8	3	11	7	4	9	8	7	11	9
20	①	★	FRESH FOOD	114	2		2	3	2			3	4	2	4	7	6	5	5	12	15	13	12	17
21	①	★	SUFFICIENT FOOD	111	1		5	7	4	10	8	6	11	16	14	3	6	2	5	3	2	2	4	2
22	①	★	COOKED FOOD	107					6	1	4	7	16	13	4	5	4	4	5	4	20	2	5	6

[**Analyze**] provides insight from the LLM, such as OpenAI (GPT), Google (Gemini), or Anthropic (Claude):

OPENAI / GPT

★★★★★

The analysis of the frequency of 'ADJ food' phrases in the COHA corpus from the 1820s to the 2010s reveals several notable trends and insights into the evolution of language surrounding food. Overall, the frequency of certain phrases demonstrates a significant increase over time, particularly from the mid-20th century onwards. For instance, "fast food" saw a dramatic rise starting in the 1970s, reflecting the cultural shift towards convenience and mass-produced food options, peaking in the 2000s. This mirrors broader societal changes related to urbanization and the fast-paced lifestyle of modern society. "Good food" has also shown a steady increase, particularly notable in the 1910s through the 2010s. This trend may suggest a growing emphasis on quality and satisfaction in food choices, perhaps influenced by the rise of culinary arts and food culture. In contrast, "bad food" maintains a more modest frequency but still reflects concerns over health and quality, especially visible in the late 20th century. Interestingly, phrases like "organic food" and "local food" began to emerge significantly in the 2000s, indicating a shift towards health consciousness and sustainability. The increasing awareness of food origins and production methods is a crucial social trend reflected in these phrases. Conversely, "daily food" and "proper food" appear to have declined in relevance, suggesting a possible shift in societal norms around what constitutes a daily diet or acceptable food standards. In summary, the data illustrates a complex interplay between cultural, health, and social factors that have influenced the language and perception of food across the decades, highlighting an evolving relationship with food that continues to shape societal values today.

GOOGLE / GEMINI

★★★★★

Based on the provided data, several interesting trends emerge regarding the use of adjectives with the word "food" in the COHA corpus from the 1820s to the 2010s. Firstly, the frequency of "good food" shows a clear upward trend, especially from the mid-19th century onwards, peaking in the 1940s and 1950s. This suggests a growing societal focus on the quality and enjoyment of food, possibly linked to increased availability and changing cultural norms around dining. Secondly, the appearance and rapid increase in frequency of "fast food" starting in the 1970s, and "Chinese food" starting in the 1950s, reflects the rise of these culinary phenomena in American society. The trends mirror the increasing prevalence and cultural impact of these food types. Conversely, terms like "wholesome food," "sufficient food," "proper food," and "necessary food" generally decline over time. This could indicate a shift away from emphasizing the basic nutritional or functional aspects of food towards more sensory or experiential qualities, or a decline in focus on simply having enough to eat. The rise of terms like "real food," "organic food," "local food," "healthy food," and "natural food" in the later decades (1980s-2010s) points to a growing awareness and concern regarding food sources, processing, and health implications. This trend reflects a contemporary focus on food origins, sustainability, and the perceived purity or authenticity of food. Finally, the data reveals the emergence and evolution of specific cuisines in the American lexicon. "Mexican food" and "Italian food" show significant growth in frequency starting in the mid-20th century, mirroring their increasing popularity and integration into the American culinary landscape.

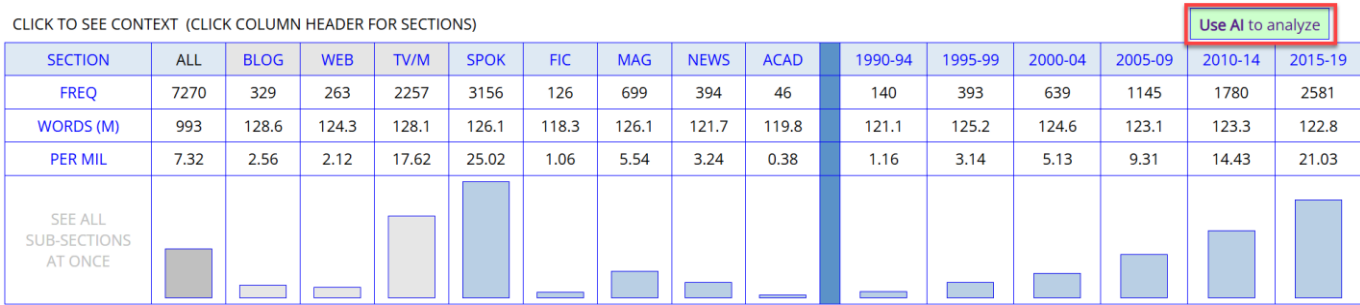
ANTHROPIC / CLAUDE

★★★★★

Looking at the data from the COHA corpus spanning the 1820s to 2010s, several interesting trends emerge in how food has been described and discussed over nearly two centuries. The most striking development is the emergence of ethnic and cultural food descriptors in the latter half of the 20th century. Terms like "Chinese food," "Mexican food," and "Italian food" were virtually non-existent before 1950 but showed significant growth thereafter. "Chinese food" in particular saw a dramatic increase from the 1960s onward, reflecting the growing popularity and acceptance of international cuisines in American culture. "Fast food" represents another notable trend, appearing for the first time in the 1970s and showing explosive growth through the 2000s, reaching its peak frequency of 2.73 in that decade. This mirrors the rise of quick-service restaurants and changing American eating habits. Some terms reflect changing social values and health consciousness. "Healthy food" was rarely used until the 2000s, but then increased dramatically to 1.27 in the 2010s. Similarly, "organic food" emerged in the late 20th century and peaked in the 2000s at 1.03. "Local food" shows a similar recent surge, reaching 0.93 in the 2010s. Certain traditional descriptors have declined over time. "Wholesome food" and "daily food" were relatively common in the 19th century but have nearly disappeared from usage. "Pure food" showed an interesting spike around 1900-1920 (possibly related to the Pure Food and Drug Act of 1906) before declining. "Good food" has remained consistently popular throughout the period, generally increasing in frequency over time, while neutral descriptors like "other food" have remained relatively stable. This suggests that while our vocabulary for describing food has expanded significantly, some basic ways of talking about food have remained constant. These trends reflect broader social changes: globalization, fast-food culture, health consciousness, and changing food values in American society.

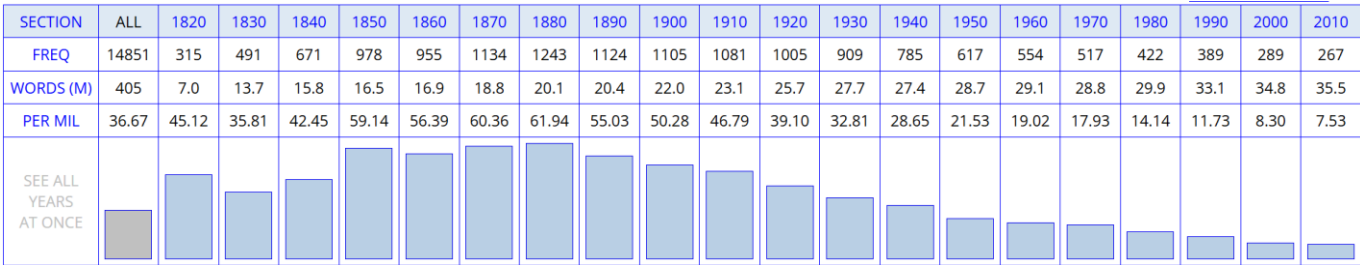
7. Analyzing differences across all sections (charts)

English-Corpora.org also shows the overall frequency of words, phrases, collocates, or syntactic constructions in the different sections of the corpora, including genres (e.g. COCA or BNC), time periods (e.g. COHA, NOW, or EEBO), or dialects (e.g. GloWbE). The LLMs provide useful analyses for the differences between sections, such as the “like construction” (CONJ PRON BE like , : and I’m like, no way) in COCA, need + NEG (e.g. needn’t, need not) in COHA, and soft day in Ireland in GloWbE.



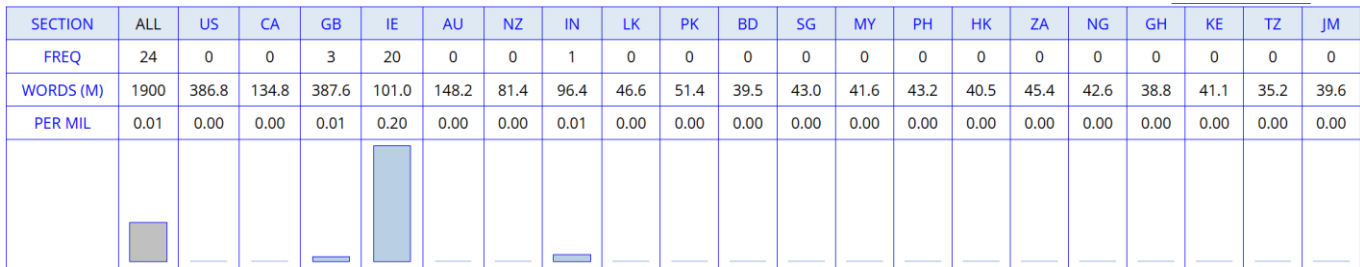
GOOGLE / GEMINI

The 'CONJ PRON BE like ,' construction is heavily favored in informal, conversational genres. Its high frequency in SPOKEN and TV/MOVIES reflects its use as a discourse marker, signaling reported speech, thought, or attitude in a casual, often performative way. MAGAZINE's moderate usage likely stems from its blend of formal and informal styles. The low frequency in ACADEMIC writing is unsurprising. Academic prose prioritizes formal, precise language and would favor constructions like "CONJ PRON stated/argued/suggested" or rephrasing altogether to avoid colloquialisms. NEWSPAPER's low usage points to a similar preference for formality and objectivity, potentially using "CONJ PRON said" instead. FICTION's low frequency could be due to authors choosing more varied narrative techniques, or using direct quotation instead of the 'like' construction. BLOG and WEB-GENL fall in the middle, reflecting their mixed content. The increasing frequency over time indicates the growing acceptance and spread of this construction in broader contexts, likely influenced by its prevalence in spoken and media forms.



ANTHROPIC / CLAUDE

The declining frequency of 'need NEG' constructions (e.g., "need not," "need n't") can be explained by the rise of competing alternative constructions, particularly "don't need to" and "don't have to." These alternatives serve the same semantic function but align better with contemporary English patterns. The data shows a clear downward trend, especially from the 1920s onward. During periods when 'need NEG' usage was high (1850s-1880s), formal writing styles dominated published texts, favoring constructions like "need not." However, as language became more colloquial throughout the 20th century, speakers increasingly preferred the more analytically transparent "don't need to" construction, which follows the regular pattern of English negation with "do-support." Additionally, the rise of "don't have to" as a common expression of negative necessity likely contributed to the decline. This construction became increasingly prevalent in the 20th century as a less formal alternative to "need not." The preference for these competing constructions reflects a broader shift in English toward more periphrastic (multi-word) expressions and away from synthetic forms like "need not." The very low frequencies in recent decades (2000s-2010s) suggest that "need NEG" has become largely restricted to formal or literary contexts, while "don't need to" and "don't have to" dominate in everyday usage.



OPENAI / GPT

The notable disparity in the frequency of the construction "soft DAY" across various English dialects can be attributed to the existence of competing expressions that are more prevalent in regions where "soft DAY" is infrequently used. In Ireland, where the usage is relatively high, the term "soft day" reflects local weather conditions—specifically, a cool, damp day with overcast skies—an expression deeply rooted in Irish culture and meteorological experiences. In contrast, countries like the United States and Australia, which exhibit a normalized frequency of 0.00, might favor alternative constructions or idiomatic expressions to describe similar weather conditions. For instance, speakers might simply use "overcast," "drizzly," or "gloomy day" to communicate the same concept. These alternatives do not carry the same cultural resonance as "soft day" does in Ireland, leading to a lack of awareness or usage of the term in these regions. Furthermore, the absence of the construction in countries like Canada and New Zealand suggests that linguistic exposure and cultural exchanges play a significant role. In these regions, weather descriptions may be influenced by local vernacular or international media, which prioritize different linguistic expressions over idiomatic local terms. Thus, the competition from other weather-related phrases, combined with varying cultural contexts, contributes to the marked differences in the frequency of "soft DAY" across dialects.

8. Analyzing concordance (Keyword in Context / KWIC) lines

KWIC (Keyword in Context, or concordance) lines are kind of the “bread and butter” of corpus analysis. But it is often difficult for language learners to analyze the many lines of text, because there are too many words that they don’t know, and because it’s hard for them to see patterns in the text. But of course that is exactly what LLMs excel at – finding patterns in data. To take a very simple example, the verb *fathom* is often preceded by a negative or doubting word or phrase: *can’t*, *couldn’t*, *cannot*, *can’t even*, *barely*, *hard to*, etc.

CLICK FOR MORE CONTEXT				TRANSLATE		ANALYZE		HELP		Use AI to find patterns	
24	2012	BLOG	...istrue.wordpress.com	trying to get you to pull over .	ca	n't	fathom	, on any level , what would keep them from pulling over			
25	1999	SPOK	ABC_Nightline	for my release . I ca n't --	ca	n't	fathom	me doing all of that time but that is my reality .			
26	2012	WEB	...ms.studentdoctor.net	to parse even my own latent intents .	ca	n't	fathom	ever getting to the bottom of another man or woman 's			
27	2012	WEB	amazon.com	a lovely breath of fresh air . I just	ca	n't	fathom	how people can dislike such an honest and moving picture , but			
28	2006	SPOK	Fox_OReilly	: There are a lot of Americans who just	ca	n't	fathom	it , and that 's why they found her not guilty .			
34	1990	NEWS	WASHPOST	, how could they not be married .	could	n't	fathom	is , so I grew up accepting granddaddy 's side of the			
35	2010	NEWS	Chicago	went back during the week , because I just	could	n't	fathom	another long wait for a table) . And several dishes sampled			
36	2004	SPOK	CBS_Morning	I -- that I heard , and I just	could	n't	fathom	how anything could be played that fast . (Footage-of-Chris-F)			
37	2014	FIC	Bk:Lowball	vanish . Why the priest cared so much Marcus	could	n't	fathom	. # Flipping the phone shut and slipping it into his chest			
38	2018	FIC	Ploughshares	dripping water . That deepened her suspicion . She	could	n't	fathom	why Gong Aihua was n't home . She had no choice but			
39	2012	WEB	starwars.wikia.com	, while still under Imperial control . Although she	could	n't	fathom	what might have drawn Vader all the way out to Kintoni ,			
40	2012	FIC	Dark-Bau	there is a reason to this effect alone . It's not magic	could	n't	fathom	. I don't know if that's what the duty is or not .			
46	2012	WEB	fuglyblog.com	child in such a dangerous situation . #	can	not	fathom	someone who would purchase such a large animal and clearly kno			
47	2012	BLOG	worldwidehippies.com	two . # For some reason , and	can	not	fathom	why , you seem to be of the opinion that a "			
48	2012	BLOG	100rsns.blogspot.com	people are STILL lining up for the J.D.	can	not	fathom	why . Law school applications have fallen off a cliff but there			
49	1997	SPOK	CBS_Sixty	know what Jett was doing . Mr-JETT :	can	not	fathom	how , in all these reports that have -- have come out			
50	2013	MAG	NewRepublic	says someone who knows them both . "	can	not	fathom	anybody in Marco ' s orbit doing anything to harm Jeb Bush			

After clicking on [Use AI to find patterns], they will see a detailed explanation of patterns like the following:

GOOGLE / GEMINI

★★★★★

Okay, here's an analysis of the patterns surrounding the word "fathom" in the provided text, focusing on aspects relevant to language learners: **1. Common Collocates (Left/Right):** *Left:* can't, cannot, hard to, difficult to*. These are overwhelmingly the most common words appearing before "fathom." *Right:* Usually a wh- word like what, why, how or a noun phrase representing something difficult to understand (the mystery, love, the pain*). **2. Semantic Prosody:** *Strongly negative.* "Fathom" overwhelmingly appears in contexts of inability to understand, disbelief, or bewilderment*. The phrases often express frustration, confusion, or the sheer difficulty of grasping something. **3. Register & Formality:** * The register is mixed, ranging from informal ("I still can't fathom why anyone would want to do this") to more formal ("Before students can fathom the mystery of faith"). However, the prevalence of contractions like "can't" suggests a generally informal or conversational tone. **4. Phraseological Patterns:** * "Can't/cannot fathom [something]" is by far the dominant pattern. This indicates a lack of understanding or an inability to comprehend. * "Hard to fathom" is another recurring phrase, expressing the difficulty of understanding. **5. Grammatical Patterns:** * "Fathom" almost always functions as a **transitive verb**, taking a direct object (the thing that is not understood). * It's frequently used in **negative constructions** (e.g., "can't fathom," "could not fathom"). * The subject of the verb is typically a person or a group of people (e.g., "I," "we," "students," "Americans"). **6. Recurring Sentence Structures:** * **[Subject] can't/cannot fathom [object/concept].** This is the core sentence structure. * Sentences often begin with a statement, followed by a "can't fathom" clause that expresses disbelief or difficulty in reconciling the statement with reality. **7. Pragmatic Function / Discourse Role:** * Expressing **disbelief, confusion, or incredulity.** * Emphasizing the **complexity or difficulty** of a situation or concept. * Conveying a sense of **frustration** or exasperation. **8. Frequency & Dispersion:** * The "can't/cannot fathom" pattern is very frequent and widely dispersed throughout the text. This is the dominant usage. **9. Thematic Context (Topics):** * **Human Behavior/Relationships:** Understanding people's actions, motivations, and feelings is a common theme. * **Politics/Social Issues:** Difficulty understanding political decisions, social phenomena, or cultural differences. * **Abstract Concepts:** Faith, love, mystery, and other abstract ideas that are hard to grasp. * **Trauma/Suffering:** Inability to comprehend extreme pain, loss, or cruelty. **Implications for Language Learners:** * **Master the "can't/cannot fathom" structure:** This is the most important takeaway. Understanding how to use this phrase will allow learners to express a common and important sentiment: the inability to understand something. * **Pay attention to the negative context:** "Fathom" almost always appears in negative contexts, so be mindful of this when using it. * **Practice using it with different objects:** Try using "can't fathom" with various nouns and concepts to expand your vocabulary and fluency. * **Be aware of the level of formality:** While generally informal, be prepared to encounter "fathom" in more formal writing as well. * **Recognize the emotional weight:** "Fathom" often conveys strong emotions like disbelief or frustration, so use it appropriately.

OPENAI / GPT

★★★★★

The analysis of the term "fathom" reveals several interesting linguistic patterns that can be beneficial for language learners and non-native speakers. **1. Common Collocates:** The word "fathom" often collocates with expressions indicating difficulty in understanding, such as "hard to fathom," "can not fathom," and "impossible to fathom." These collocates suggest that the term is frequently used in contexts of confusion or incomprehensibility. **2. Semantic Prosody:** The semantic prosody surrounding "fathom" leans towards a negative or challenging connotation. Most instances convey a sense of struggle to understand or grapple with complex emotions, situations, or societal issues. **3. Register & Formality:** The language surrounding "fathom" ranges from informal to semi-formal. Phrases like "I can't fathom" appear in both casual discussions and more serious contexts, indicating that the term is versatile across different registers. **4. Phraseological Patterns:** Common phrases include "can not fathom," "hard to fathom," and "impossible to fathom," suggesting a pattern of expressing bewilderment or shock. **5. Grammatical Patterns:** "Fathom" is primarily used as a verb and appears in various grammatical structures, including modal constructions ("can not fathom") and passive constructions, emphasizing its role in expressing subjectivity and personal struggle. **6. Recurring Sentence Structures:** The term often appears in declarative sentences that express personal reflection or societal critique, indicating its role in conveying individual experiences and collective sentiments. **7. Pragmatic Function / Discourse Role:** "Fathom" serves to express attitudes of confusion, disbelief, or emotional distress. It can also emphasize the speaker's struggle to comprehend a situation, adding depth to the discourse. **8. Frequency & Dispersion:** The term is used frequently throughout the text, indicating that it is a common expression for discussing complex thoughts or feelings, and it

9. Generating word and phrase lists

As was mentioned previously, when LLMs try to generate word and phrase lists by themselves, these lists don't agree very well with actual corpus data. But at English-Corpora.org, we *combine* the LLM suggestions and the corpus data. The LLM suggests what it *thinks* are the most relevant words and phrases for to a particular topic, and then we rank those suggestions, using the actual frequency in the corpus. Here are just a few examples, and note that for reasons of space, this page shows just the first 10-11 entries for each search, while the corpus can show up to 200 words and phrases.

[NOW corpus] Prompt: [find words and phrases related to] **tariffs**

OPENAI / GPT Generate more (Show query sent to AI) ★★★★★

ON CLICK:

CONTEXT

TRANSLATE (ES)

ENTIRE PAGE

GOOGLE

IMAGE

PRON/VIDEO

BOOK

THESAURUS

 (HELP)

SAVE	ⓘ	★	WORD / PHRASE (RE-USE)	FREQ	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
1	ⓘ	★	tariffs	618020	984	1342	1981	2783	2413	3227	12590	15499	84855	79102	28810	14793	16999	12403	55994	284245
2	ⓘ	★	competitive advantage	56574	623	868	1050	1292	1686	1764	4003	4337	3898	5477	7274	7680	6167	4327	3956	2172
3	ⓘ	★	consumer prices	35253	176	224	202	232	295	557	2142	2200	1792	1706	1528	4822	9267	4750	3389	1971
4	ⓘ	★	protectionism	32231	301	314	322	299	295	296	2978	6371	5799	3834	2478	1211	1313	1265	2352	2803
5	ⓘ	★	economic sanctions	28790	107	158	363	272	649	721	1682	2272	2804	3309	2212	2395	8260	1670	1138	778
6	ⓘ	★	revenue generation	26603	115	183	253	321	312	341	2506	2481	2346	3049	2421	2475	2802	2611	2806	1581
7	ⓘ	★	World Trade Organization	23387	164	226	293	305	341	391	1238	1744	2675	2786	3144	3120	2196	1653	1414	1697
8	ⓘ	★	global supply chains	18761	15	45	51	70	75	97	296	410	619	736	3397	2924	3997	2100	1819	2110
9	ⓘ	★	local markets	17943	177	232	329	370	399	455	1404	1419	1729	1704	2048	1646	1816	1408	1765	1042
10	ⓘ	★	trade barriers	17352	75	133	196	233	238	297	1221	1328	2265	1358	1097	930	1018	1018	1588	4357

[COHA corpus] Prompt: [find words and phrases related to] **jobs that were done mainly by women in the 1800s**

SAVE	ⓘ	★	WORD / PHRASE (RE-USE)	FREQ	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
1	ⓘ	★	Cook	21759	70	305	314	330	488	653	790	751	913	1109	1278	1202	998	1443	1049	1534	1422	1962	2727	2421
2	ⓘ	★	Housekeeper	3630	23	59	159	194	332	266	246	173	254	274	220	166	173	156	132	161	131	148	209	154
3	ⓘ	★	Governess	1462	20	23	15	97	135	151	100	210	54	72	77	164	72	62	36	45	28	33	37	31
4	ⓘ	★	Nanny	1178	6		2	6	4	3	9	3	7	202	13	76	26	32	28	38	42	139	344	198
5	ⓘ	★	Midwife	771		1	3	3	6	1	5	42	5	6	12	38	53	37	47	42	86	73	114	197
6	ⓘ	★	Dressmaker	666		3	3	23	24	46	28	30	67	97	39	70	78	29	23	20	26	16	23	21
7	ⓘ	★	Seamstress	617	2	26	10	50	53	44	25	54	36	45	16	26	17	34	21	22	14	42	39	41
8	ⓘ	★	Chambermaid	584		18	35	81	39	33	45	27	27	29	39	40	42	21	22	16	22	21	16	11
9	ⓘ	★	Milliner	460	11	16	41	59	49	40	35	34	30	38	30	13	22	3	7	2	4	2	6	18
10	ⓘ	★	Spinner	313	2	10	5	8	10	11	21	10	11	10	12	18	5	15	17	11	17	45	38	37
11	ⓘ	★	Nursemaid	289					5	3	9	3	6	9	40	33	22	19	23	18	27	15	26	31

And they can even be more “abstract”, such as: [COCA corpus] Prompt: [find words and phrases related to] **ways to express condolences**

SAVE	ⓘ	★	WORD / PHRASE (RE-USE)	FREQ	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
1	ⓘ	★	thinking of you	1289	140	101	298	56	122	25	19	4	84	110	94	94	78	64
2	ⓘ	★	in sympathy	760	23	37	6	29	185	46	26	58	77	70	67	52	37	47
3	ⓘ	★	our thoughts and prayers	497	50	45	9	109	2	20	61		5	12	15	43	34	92
4	ⓘ	★	in remembrance	446	35	95	22	18	31	27	37	23	17	26	41	30	15	29
5	ⓘ	★	with sympathy	400	23	25	6	11	103	18	14	24	40	30	34	29	22	21
6	ⓘ	★	in our thoughts	241	37	30	13	31	6	13	19	5	13	13	16	15	10	20
7	ⓘ	★	deepest sympathies	196	11	11	31	22	5	10	19		8	11	15	11	7	35
8	ⓘ	★	we are so sorry	133	3	6	40	13	7		2		4	2	12	10	16	18
9	ⓘ	★	in loving memory	119	9	4	12	13	3	11	12	2	5	11	8	14	9	6
10	ⓘ	★	heartfelt condolences	82	15	3	4	9	1	5	13		2		3	9	2	16

It is also possible to add lists of “good” and “bad” examples of words and phrases. For example, if we want to search in the NOW corpus for Japanese automobile *models*, but it is giving us words and phrases like *Toyota*, *Honda*, *fuel efficiency*, or *electric vehicles*, then we can add those to a “bad examples” list, and add *Honda Civic*, *Toyota Prius*, and others to a “good examples” list. The LLM will then generate a better list, such as the following:

S <small>AVE</small>	①	★	WORD / PHRASE (RE-USE)	FREQ	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
1	①	★	Honda Civic	16395	350	397	407	288	262	351	1195	1154	1123	1903	1853	1615	2002	1515	1271	709
2	①	★	Toyota Corolla	15299	223	232	268	347	346	325	1034	1121	1323	1830	1424	1319	1603	1642	1537	725
3	①	★	Toyota Camry	10912	124	158	217	185	217	201	724	1056	830	1080	1039	1071	1159	1083	1156	612
4	①	★	Nissan Leaf	8123	191	323	268	245	192	160	371	708	864	1121	739	720	863	629	454	275
5	①	★	Honda Accord	7516	125	186	213	198	156	178	510	621	473	919	773	787	927	556	567	327
6	①	★	Toyota RAV4	6127	53	46	82	79	82	95	168	203	223	587	810	830	767	680	738	684
7	①	★	Suzuki Swift	5615	78	158	135	71	51	71	194	331	506	553	399	272	598	1049	775	374
8	①	★	Toyota Prius	5494	245	269	247	188	147	158	359	364	348	469	538	474	587	475	369	257
9	①	★	Toyota Land Cruiser	4733	31	30	54	83	66	78	283	375	444	458	441	532	582	484	452	340
10	①	★	Honda CR-V	4728	71	120	90	88	87	81	177	331	344	485	582	435	490	561	541	245
11	①	★	Mazda3	4414	112	173	199	222	275	146	292	241	297	887	640	423	162	155	114	76

Likewise, in the EEBO corpus (1470s-1690s), if we search for “terms for sicknesses, which were common in the 1500s-1600s”, but it is not giving us what we want, we could add “good examples” like *putrid fever*, *grippe*, *ague*, *French pox*, *melancholia*, and *palsy* (and perhaps spellings from the 1500s-1600s), we might then see a list like the following:

WORD / PHRASE (RE-USE)	FREQ	1470s	1480s	1490s	1500s	1510s	1520s	1530s	1540s	1550s	1560s	1570s	1580s	1590s	1600s	1610s	1620s	1630s	1640s	1650s	1660s
pestilence	10310	11	67	33	12	2	21	176	155	376	421	585	586	274	979	695	660	817	536	1043	963
consumption	5742						3	6	13	11	42	138	172	130	191	362	221	433	307	752	733
flux	5051		12	2	2	1		2	21	53	5	23	78	61	91	201	64	480	115	965	588
ague	4363		3	2	2	1	22	4	21	52	100	273	150	201	314	290	215	374	247	587	345
scurvy	2241														1	1	4	28	44	274	156
lethargy	1238							1		1			5	4	12	30	18	73	46	216	144
jaundice	732												10		4			18	49	228	40
dropsy	597		1				42	9	13	38	13	60	6	14	6	12	9	48	11	102	30
palsy	448			2			3	5	9	33	9	22	2	3	9	18	7	20	14	72	47
erysipelas	420								1		22	9	120		1	4		52		25	35
falling sickness	407															1		2	15	153	50
dysentery	398										1					4		34	5	57	63

It is also possible to see a short description (generated by the LLM) of the words and phrases, such as the following in the iWeb corpus for the prompt: [find words and phrases related to] nuclear energy.

S <small>AVE</small>	①	★	WORD / PHRASE (RE-USE)	DEFINITION	FREQ
1	①	★	Coolant	Fluid removing heat from reactor core	78825
2	①	★	Enrichment	Increasing fissile isotope concentration in nuclear fuel	55312
3	①	★	Moderator	Material slowing neutrons to sustain fission	45638
4	①	★	Half-life	Time for half of radioactive material to decay	24564
5	①	★	Fukushima	Site of nuclear disaster following 2011 tsunami	15054
6	①	★	Chain reaction	Self-sustaining series of fission reactions	12907
7	①	★	Chernobyl	Site of catastrophic nuclear accident in 1986	11070
8	①	★	Decommissioning	Process of safely closing a nuclear reactor	10100
9	①	★	Nuclear power plant	Facility generating electricity using nuclear reactors	9533
10	①	★	Nuclear waste	Byproducts of nuclear reactions requiring disposal	6637
11	①	★	Reprocessing	Recycling spent nuclear fuel for reuse	6407

Finally, something that should be useful for language learners is the ability to enter a word or phrase in their native language. It will then translate this to English (sometimes in several different ways) and then find matching words and phrases from English. For example, a Spanish speaker could enter the following simple two word prompt in COCA: [find words and phrases related to] **Spanish “manguera”** (where *manguera* means “garden hose”), and with that simple two word prompt, they would see the following (which also includes a short definition, as in the example above).

SAVE	①	★	WORD / PHRASE (RE-USE)	DEFINITION		FREQ	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
1	①	★	hose	flexible tube for conveying liquids or gases.	▶	5563	400	377	947	335	1729	1220	386	169
2	①	★	plumbing	system of pipes for water supply and drainage.	▶	4246	566	446	580	274	570	907	636	267
3	①	★	water source	origin from which water is obtained.	▶	541	54	72	44	16	21	131	75	128
4	①	★	garden hose	a hose used for watering plants and gardens.	▶	500	30	45	55	40	127	132	61	10
5	①	★	fire hose	durable hose used by firefighters to deliver water.	▶	291	27	19	58	30	53	51	40	13
6	①	★	irrigation system	method for supplying water to crops or gardens.	▶	283	18	21	17	6	16	58	35	112
7	①	★	drip irrigation	efficient watering method delivering water directly to roots.	▶	175	7	12	3	5	1	86	14	47
8	①	★	rubber hose	flexible tube made from rubber for various uses.	▶	98	2	8	20	8	32	17	8	3
9	①	★	pressure washer	high-pressure device for cleaning surfaces with water.	▶	64	4	4	6	2	3	34	6	5
10	①	★	leak detection	process of identifying unwanted water escape in systems.	▶	47	3	2		2		8	10	22
11	①	★	hose clamp	device used to secure hoses onto fittings.	▶	37	8	3	2		2	21		1

Or a German speaker could enter the following prompt in the 325 million word TV Corpus: [find words and phrases related to] **the German phrase "Es tut mir Leid"** (which roughly means “I’m sorry”), and they would see the following phrases. Not all of these are completely relevant, but it’s a good start, and it’s definitely not anything that other corpora could provide via their search syntax.

SAVE	①	★	WORD / PHRASE (RE-USE)		FREQ	1950s	1960s	1970s	1980s	1990s	2000s	2010s	US/CA	UK/IE	AU/NZ	Misc
1	①	★	I'm sorry	▶	373010	951	3870	4401	8251	17072	52267	99693	159441	23323	2808	933
2	①	★	I did n't mean to	▶	11926	40	141	188	319	673	1743	2859	5030	774	125	34
3	①	★	my apologies	▶	3130	10	50	60	85	142	303	915	1275	250	22	18
4	①	★	it was my fault	▶	2770	7	39	38	71	156	406	668	1104	241	33	7
5	①	★	please forgive me	▶	1320	4	32	18	61	67	152	326	537	102	13	8
6	①	★	I'm truly sorry	▶	604		9	9	12	31	77	164	245	50	4	3
7	①	★	I take full responsibility	▶	514		2	5	9	17	82	142	220	31	6	
8	①	★	I regret that	▶	456	2	14	6	8	34	41	123	194	33	1	
9	①	★	I regret to inform you	▶	204	1	4	1	4	15	28	49	84	18		
10	①	★	I hope you can forgive me	▶	186				5	12	18	58	85	7	1	
11	①	★	I am deeply sorry	▶	104		1	1		2	13	35	38	9		5
12	①	★	I feel bad about that	▶	68					4	10	20	32	1	1	

English-Corpora.org already provided users with a much larger range of searches than any other online corpus site, with [searches](#) by word form, lemmas, [variable length](#) queries, [collocates](#), [topics](#) (words co-occurring anywhere in the text), [part of speech](#), synonyms (for example [=beautiful](#) or [=strong ARGUMENT](#)), [customized word lists](#), and (with [browse](#)) by word frequency range, definition and even by pronunciation.

The use of AI-generated words and phrases now takes this to an entirely new level. Users can search by any “concept”, as shown by the searches above, including concepts in *another language*. And as AI/LLMs improve even more, it will create even more powerful searches of the data in the corpora at English-Corpora.org.