

Showing, limiting, and sorting by the number of texts

In addition to seeing the frequency of words and phrases, you can now see the number of texts in which these words and phrases occur, and sort and limit the results by the number of texts.

But **why** would you want to do this? Well, sometimes there are **words or phrases** that are **limited to just a few texts** in the corpus. If those texts weren't in the corpus, the frequency of the word or phrase might be much lower, or it might not occur at all.

For example, consider these results for the search *cold NOUN* in the [COHA corpus](#): Corpus of Historical American English. (Note that the screenshots here come from the LIST view, but you can also see the number of distinct texts in the CHART view, including the number of texts by sub-genre, year, etc.)

HELP	ALL FORMS (SAMPLE): 100 200 500	TEXTS	FREQ	TOTAL 30,153 UNIQUE 3,568 +
1	COLD WATER	1800	2933	
2	COLD WAR	1290	2135	
3	COLD WEATHER	854	1083	
4	COLD AIR	710	910	
24	COLD MOUNTAIN	57	152	
25	COLD SHOWER	130	151	
26	COLD MEAT	125	149	
27	COLD STONE	132	142	
28	COLD BATH	92	135	
29	COLD HARBOR	55	135	

When you think of nouns that occur with *cold*, you might think of *water*, *weather*, *shower*, or *bath*, but probably not *mountain* or *harbor*, which are used mainly as place names in a limited number of texts.

This is often seen best in **comparisons between one section** of the corpus and another. For example, the following are the results for *old NOUN* in COHA, for the 1970s-2010s (left) and the 1840s-1890s (right). Notice that *Old Hurricane* only appears in two texts, and *Old Sophy* and *Old Mirabel* appear in one text each. (These are tagged as NOUN, but they probably should have been tagged as NAME: proper noun.)

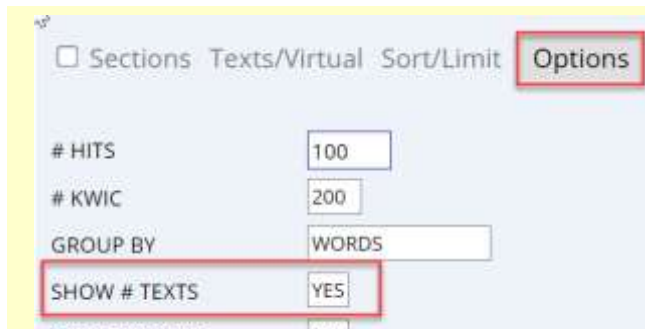
SEC 1 (1970, 1980, 1990, 2000, 2010): 162,104,741 WORDS						SEC 2 (1840, 1850, 1860, 1870, 188...): 108,562,065 WORDS					
WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1 OLD GUY	391 (24%)	0 (0%)	2.4	0.0	241.2	1 OLD HURRICANE	295 (11%)	1 (1%)	2.7	0.0	440.5
2 OLD BUDDY	246 (15%)	0 (0%)	1.5	0.0	151.8	2 OLD SINNER	47 (17%)	1 (1%)	0.4	0.0	70.2
3 OLD MOVIES	154 (9%)	0 (0%)	1.0	0.0	95.0	3 OLD SOPHY	74 (27%)	0 (0%)	0.7	0.0	68.2
4 OLD BASTARD	135 (8%)	0 (0%)	0.8	0.0	83.3	4 OLD CLERGYMAN	45 (16%)	1 (1%)	0.4	0.0	67.2
5 OLD MOVIE	115 (7%)	0 (0%)	0.7	0.0	70.9	5 OLD SCENES	40 (14%)	1 (1%)	0.4	0.0	59.7
6 OLD BOYFRIEND	94 (6%)	0 (0%)	0.6	0.0	58.0	6 OLD NEGRESS	60 (22%)	0 (0%)	0.6	0.0	55.3
7 OLD PALS	79 (5%)	1 (1%)	0.5	0.0	52.9	7 OLD MIRABEL	57 (21%)	0 (0%)	0.5	0.0	52.5

And if we looked at the concordance lines for these phrases, they occur as the names of characters in just the one or two texts, as with *Old Hurricane*:


21	1867 FIC HiddenHand	Q	you suspect? "The woman nodded. "It was --" said Old Hurricane , scooping and whispering a name that was near to no one but the
22	1867 FIC HiddenHand	Q	wine, medicine, clothing and every comfort that your condition requires," said Old Hurricane , rising and calling in the clergyman, with whom he soon after left
23	1867 FIC HiddenHand	Q	, for now I shall have the game in my own hands!" muttered Old Hurricane to himself. "Ah! Gabrielle Le Noir, better you had cast
24	1867 FIC HiddenHand	Q	master's outer garments. But, let his family wonder as they would, Old Hurricane kept his own counsel -- only just as he was going away, lest
25	1867 FIC HiddenHand	Q	head. After which undeniable apothegm the conversation came to a stand. Meanwhile, Old Hurricane pursued his journey -- a lumbering, old-fashioned stage-coach rid
26	1867 FIC HiddenHand	Q	and porters. At length, taking up his heavy carpet-bag in both hands, Old Hurricane began to lay about him, with such effect that he speedily cleared a

In other words, if the corpus didn't have those few texts, the phrases might not occur in the corpus at all. But for phrases like *old sinner* (47 tokens in 35 texts) and *old scenes* (40 tokens in 37 texts), the phrases are spread out more through the entire corpus.

How to do it



It is very easy to see the number of texts. In the search form, just click on [Options] and then set [Show # texts] to [YES].



You can also sort the results by the number of texts in which they occur (use [SORTING] in the search interface).

Perhaps more importantly, you can **limit the results** to just those words or phrases that occur in at least a **certain number of texts**. For example, if we set the minimum number of texts to [5], then the results would be the following (notice that *Old Hurricane*, *Old Sophy*, and *Old Mirabel* are no longer there, since they did not occur in at least five texts):

SEC 1 (1970, 1980, 1990, 2000, 2010): 162,104,741 WORDS						SEC 2 (1840, 1850, 1860, 1870, 188...): 108,562,065 WORDS							
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	OLD GUY	391 (100%)	0 (0%)	2.4	0.0	241.2	1	OLD SINNER	47 (100%)	1 (100%)	0.4	0.0	70.2
2	OLD BUDDY	246 (100%)	0 (0%)	1.5	0.0	151.8	2	OLD CLERGYMAN	45 (100%)	1 (100%)	0.4	0.0	67.2
3	OLD MOVIES	154 (100%)	0 (0%)	1.0	0.0	95.0	3	OLD SCENES	40 (100%)	1 (100%)	0.4	0.0	59.7

In the results (such as for words with **clean**, below), the number of texts (such as 315 for *McLean*) is highlighted in red when it is found in a limited number of texts. This is perhaps most useful when we have chosen to see the **results by section** in the search interface.



HELP	★	TEXTS	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	
11	★	LLBLANDE	555	659	28 (13)	30 (11)	31 (22)	30 (20)	31 (25)	23	35 (27)	27	26	20	37	14	
12	★	MCLEAN	315	653	1	20 (6)	6	26	14 (7)	9 (4)	28 (14)	12	13	21	46	48	42
13	★	CLEANSSED	535	646	12 (11)	29 (26)	33 (26)	30 (21)	34 (30)	29 (22)	46 (33)	30 (23)	31 (23)	25 (24)	46 (34)	34	26 (23)
14	★	MACLEAN	125	522		2		33 (2)		8 (4)	7 (3)	138 (5)	2	6 (3)	25 (7)	14 (7)	
15	★	CLEAN-CUT	404	475			2 (1)	1 (1)		5 (3)	4	18	34 (27)	44 (43)	62 (44)	35 (30)	37 (31)

Notice that the 33 tokens of *MacLean* occur in only two texts in the 1860s and only four texts for the 138 tokens in the 1900s. This is pretty good evidence that this is the name of someone who is the focus of discussion in these handful of texts, rather than a word that is spread better throughout the section (such as *cleansed* or *clean-cut*).

CONCORDANCE DISPLAY

You can also see information on the number of texts as part of the concordance display (also known as the Keyword in Context display, or KWIC), even if you haven't chosen to see this in the frequency display. For example:

FIND SAMPLE: 100 200
PAGE: 1 / 4

352 ENTRIES: 70 TEXTS
LIMITS: NONE
SORTING: YEAR, GENRE

CLICK FOR MORE CONTEXT				SAVE	TRANSLATE	ANALYZE	HELP
1	1828	FIC	AndrewJackson	🔍	🔍	🔍	general. Jackson-man. A Jackson-man What! (with vehemance and emphasis,) Old Hickory a murderer -- why, consider Turncoat, (pulling Turncoat
2	1828	FIC	AndrewJackson	🔍	🔍	🔍	of a man's head, is enough to make a saint swear -- call Old Hickory a murderer, who has saved the lives of millions, is downright slanderation
3	1828	FIC	AndrewJackson	🔍	🔍	🔍	, who instead of stinging him, fell by the Kentuck rifles. They say Old Hickory can not spell rifle, but John Bull as well as Uncle Sam,
4	1828	FIC	AndrewJackson	🔍	🔍	🔍	Jackson-man. A Jackson-man Yes, the martial law, so say no more about Old Hickory and murder. Why, man you mought, for the varsal world,
5	1828	FIC	AndrewJackson	🔍	🔍	🔍	Clay, the Scratchitary of State, I think they say he is, calls Old Hickory . -- But I was saying, as how, you mought as well
6	1828	FIC	AndrewJackson	🔍	🔍	🔍	I tell you the honest revarnished truth, and not a word more -- that Old Hickory had no more hand in the construction of the six miluntary men,
7	1828	FIC	AndrewJackson	🔍	🔍	🔍	for Jackson! sixty death-warrants! aha! Jackson is no murderer! Huzza for Old Hickory ! huzza! huzza! Huzza! Enter a Subaltern. Subalter

Finally, you can click on SHOW TEXT ID, which will indicate which entries are from the same text as the previous line (and which are highlighted in red):

FIND SAMPLE: 100 200
PAGE: 1 / 4

352 ENTRIES: 70 TEXTS
LIMITS: NONE
SORTING: YEAR, GENRE

CLICK FOR MORE CONTEXT				SAVE	TRANSLATE	ANALYZE	HELP
1	8566	1828	FIC	AndrewJackson	🔍	🔍	general. Jackson-man. A Jackson-man What! (with vehemance and emphasis,) Old Hickory a murderer -- why, consider Turncoat, (pullir
2	8566	1828	FIC	AndrewJackson	🔍	🔍	of a man's head, is enough to make a saint swear -- call Old Hickory a murderer, who has saved the lives of millions, is downright slanc
3	8566	1828	FIC	AndrewJackson	🔍	🔍	, who instead of stinging him, fell by the Kentuck rifles. They say Old Hickory can not spell rifle, but John Bull as well as Uncle Sam,
4	8566	1828	FIC	AndrewJackson	🔍	🔍	Jackson-man. A Jackson-man Yes, the martial law, so say no more about Old Hickory and murder. Why, man you mought, for the varsa
5	8566	1828	FIC	AndrewJackson	🔍	🔍	Clay, the Scratchitary of State, I think they say he is, calls Old Hickory . -- But I was saying, as how, you mought as well
6	8566	1828	FIC	AndrewJackson	🔍	🔍	I tell you the honest revarnished truth, and not a word more -- that Old Hickory had no more hand in the construction of the six miluntary men,
7	8566	1828	FIC	AndrewJackson	🔍	🔍	for Jackson! sixty death-warrants! aha! Jackson is no murderer! Huzza for Old Hickory ! huzza! huzza! Huzza! Enter a Subaltern. Subalter
8	7176	1832	FIC	SwallowBarnASojourn	🔍	🔍	he, with particular emphasis on the last word. " Do you know what old Hickory said down there in the Creek nation, in the war, when t

All of these features will make it easier for you to see which words and phrases are limited to just a few texts, and which ones are spread more evenly throughout the entire corpus.