COCA_ReadMe

The corpora we will most often use are here: http://corpus.byu.edu/ .**You must register for the site. It is free.**

The corpus on this page we will most often use is the Corpus of Contemporary American English. This corpus is referred to as the COCA. You can link to the corpus here: http://corpus.byu.edu/coca/

This corpus is approximately 450-million words and includes texts from 1990-2012. Each year Mark Davies and BYU add 20 million words to the corpus. Thus, they call the COCA a monitor corpus because it is designed to allow researchers to investigate language change over time.

The same search terms and search syntax can generally be used in all corpora.

USING THE CORPUS

Search types

1. **exact word**: a search for an exact word, e.g. beautiful.

2. **phrases:** a search for an exact phrase, e.g. beautiful sunset

3. **lemmas:** a search for all forms of a word, e.g. search for [study] will give results for study, studies, studying, etc. For a lemma search, the word must be in [ ] .

4. **wildcard:** a search that allows you to find different forms of a word. A search for beaut* will give results beautiful, beautification, beauty, etc.

    a. A wildcard * can be place at any place in the word and can be for any # of letters.

        un*ly will give results unlikely, unusually, etc.

    b. A wildcard ? is only for one letter.

        s?ng will give results sing, sang, sung, song, etc.

5. **synonym**: allows you to find all synonyms of a word. For example, [=beautiful] will yield beautiful, attractive, gorgeous, etc.

6. **part of speech** (pos): allows you to find an exact part of speech.

    a. a search for [np*] will give a list of the most frequent proper nouns in the corpus

    b. a search of [j*] man will give you a list of the most frequent adjective + man collocations

    c. a search for study.[v*] will only give results for 'study' used as a verb

The pos options are listed in a dropdown menu on the left side of the interface.

7. **multiple words**: a search for talk|speak|lecture will give you results for all three words

NOTE 1: These are only a few of the search syntax possibilities for the corpus. There are more options listed on the "search syntax" page on the COCA.

NOTE 2: You can combine different types of searches. E.g., you can have search strings to help extract more specific findings.

NOTE 3: Errors with search syntax will occur. When you have a problem, consult the help page or email me.

The Interface

1. LIST: provides the most basic results.

2. CHART: provides results across the 5 major registers in a bar table format

Image 2: Chart of 'beautiful'

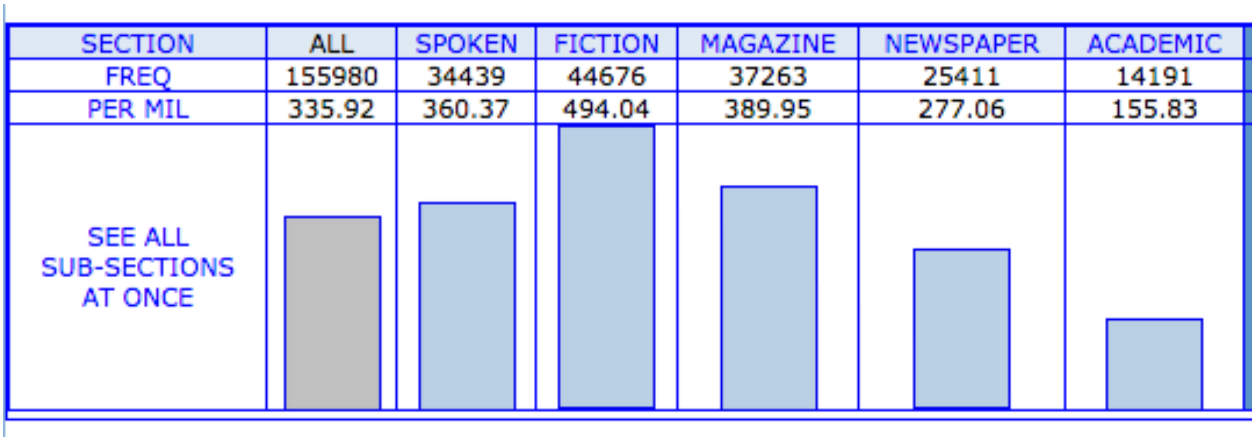
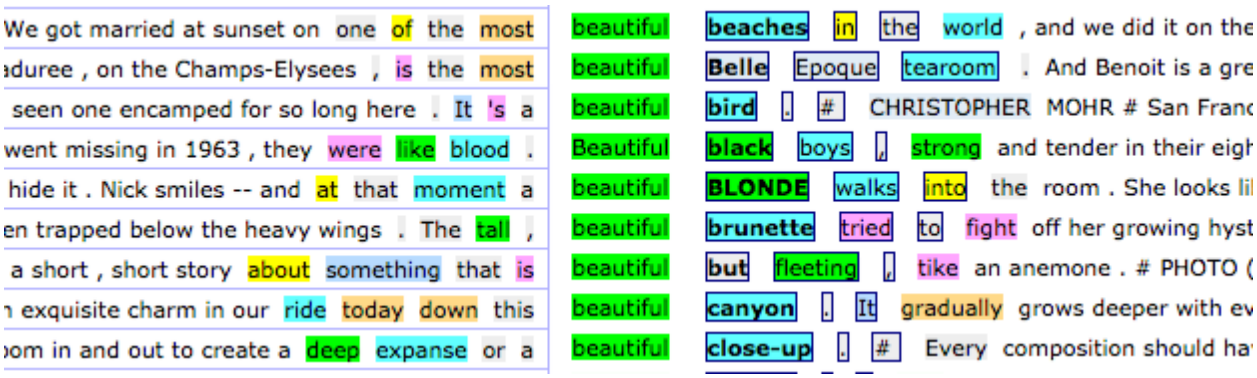


Image 2 shows the results if you were to search for 'beautiful' using the chart option. The chart shows the frequency of usage across the 5 main registers in the corpus. FREQ is the raw frequency; this is the total number of times the word occurs in the corpus for that particular section. PER MILL means this is how many times the word occurs for every 1 million words in the register.

3. KWIC (key word in context): provides results with the different word classes color coded

Image 2: KWIC



We got married at sunset on one of the most	beautiful	beaches in the world , and we did it on the
aduree , on the Champs-Elysees , is the most	beautiful	Belle Epoque tearoom . And Benoit is a gre
seen one encamped for so long here . It 's a	beautiful	bird . # CHRISTOPHER MOHR # San Franc
went missing in 1963 , they were like blood .	Beautiful	black boys , strong and tender in their eigh
hide it . Nick smiles -- and at that moment a	beautiful	BLONDE walks into the room . She looks li
en trapped below the heavy wings . The tall ,	beautiful	brunette tried to fight off her growing hyst
a short , short story about something that is	beautiful	but fleeting , like an anemone . # PHOTO (
n exquisite charm in our ride today down this	beautiful	canyon . It gradually grows deeper with ev
oom in and out to create a deep expanse or a	beautiful	close-up . # Every composition should ha

Image 3 shows the adjective "beautiful" in green in the center of what are called concordance lines. Concordance lines are the lines of output with your search word in the center. We can see from the KWIC lines that "beautiful" is commonly followed by a noun.

4. COMPARE: provides results comparing the collocates of two words

Image 3: Compare

WORD 1 (W1): **BEAUTIFUL** (3.76)

	WORD	W1	W2	W1/W2	SCORE
1	SONG	178	1	178.0	47.4
2	MOUNTAINS	86	0	172.0	45.8
3	PRINCESS	83	0	166.0	44.2
4	ABSOLUTELY	213	2	106.5	28.4
5	COUNTRYSIDE	50	0	100.0	26.6
6	SKY	94	1	94.0	25.0
7	HORSE	45	0	90.0	24.0
8	SUNNY	85	1	85.0	22.6
9	HAUNTINGLY	42	0	84.0	22.4
10	BREATHTAKINGLY	40	0	80.0	21.3

WORD 2 (W2): **ATTRACTIVE** (0.27)

	WORD	W2	W1	W2/W1	SCORE
1	INVESTORS	76	0	152.0	570.8
2	INVESTMENT	74	0	148.0	555.7
3	ALTERNATIVE	138	1	138.0	518.2
4	OPTION	136	1	136.0	510.7
5	BUYERS	39	0	78.0	292.9
6	ALTERNATIVES	29	0	58.0	217.8
7	RATES	28	0	56.0	210.3
8	STOCKS	50	1	50.0	187.8
9	CONSUMERS	25	0	50.0	187.8
10	POLITICALLY	25	0	50.0	187.8

Image 3 shows the words that collocate with 'beautiful' and 'attractive'. While these two items may be considered near-synonyms, they have distinct semantic fields. This means they are used with different words, in different contexts, to produce different meanings.

Searching

1. Enter the item you want to search for in the search bar.

2. If you are looking for what occurs with the word, choose "collocates".

Collocates are words that frequently occur together, e.g. *strong coffee*, not *powerful coffee*.

You can adjust the number of words to the left and right for your search. For example, if you want only to know the words that occur exactly 1 word before your search item, select 1 in the left and 0 in the right. This will produce only a list of the words which occur before your word.

3. **POS List:** Choosing an item from the parts-of-speech list (POS) list lets you be more selective with your results. You can put the POS tag before or after your search word, and you can use more than one POS tag.

4. The next section allows you to compare different registers, years, etc. For example, you can compare the use of the adverb "perhaps" between spoken and academic English.

The results below display the findings for a search that compares the use of "however" between spoken and academic registers.

Image 4: Comparing Registers

SEC 1: 95,565,075 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	HOWEVER	9343	81925	97.77	899.62	0.11

Tokens 1: This is the number of times "however" was used in spoken English.

Tokens 2: This is the number of times "however" was used in academic.

PM 1: PM stands for parts per million. This means "however' is used 97 times per million words of spoken English.

PM2: This means "however' is used 899 times per million words of academic English.

5. Sorting: I advise you to use the default "sort by: frequency".

To the Corpus>>>