О чем речь

Корпус современного американского английского (Corpus of Contemporary American English) — часть проекта Brigham Young University, содержащего также корпуса разных вариантов английского языка (в основном) и несколько корпусов романских языков, а также позволяющего загружать некоторые статистические данные по всему корпусу в целом, без выполнения запросов.

В нем собраны тексты устной (10%) и письменной (90%) речи в период с 1995 по 2015 год, распределенные по жанрам следующим образом:

GENRE	COCA (millions of words)
Spoken	109
Fiction	105
Popular magazines	110
Newspaper	106
Academic	103

Он регулярно обновляется и является одним из самых используемых корпусов английского языка, в т.ч. среди всех корпусов BYU: примерно 65,000 уникальных пользователей в месяц.

Онлайн-интерфейс корпуса доступен по адресу corpus.byu.edu/coca/

Посетителям предоставляется возможность нескольких поисковых запросов, после чего на сайте становится необходимо зарегистрироваться для продолжения работы с поиском. Зарегистрироваться можно в качестве пользователей/исследователей разного уровня, от чего позже зависит ограничение на поисковые запросы и объем данных, с которыми можно работать.

Корпус предоставляет возможность выгрузить в офлайн определенное количество полных текстов и статистических данных; полный объем этих данных можно загрузить за

определенную плату. Согласно информации на сайте проекта, такие ограничения вызваны нагрузками на сервер.

Дизайн

Пастельные цвета, жирный белый шрифт и белые иконки хорошо читаются и выделяются на голубом фоне кнопок в верхней строке. Остальные кнопки, находящиеся рядом со строкой поиска, выполнены в голубом или салатовом цвете и обведены темно-синей рамкой, поэтому хорошо выделяются на белом фоне.

Цветовое решение во всех вкладках и на всех страницах одинаковое – пастельные оттенки синего и голубого на белом фоне. Такие цвета не привлекают внимания и не отвлекают от поиска и обработки информации.

На всём сайте выделяются ярко-желтая иконка входа в учётную запись и красная кнопка с белым восклицательным знаком, если пользователь находится в мобильной версии сайта.

Плюс – наличие не только десктопной, но и мобильной версии сайта.

Минус – мобильная версия сайта не подстраивается под горизонтальное положение экрана. Масштаб остаётся тем же, что и при вертикальном просмотре, из-за чего возникает неудобство при чтении. При этом на сайте всплывает предупреждение и совет держать телефон в вертикальном положении.

Пользователю, зашедшему в первый раз, может быть сложно сразу догадаться о значении иконок в верхней строке, поэтому во вкладке help есть возможность посмотреть легенду.

Строка поиска расположена на видном месте, сразу бросается в глаза, что удобно для пользователей. Результаты поиска можно переключать по вкладкам внутри страницы, что позволяет параллельно работать с данными о частоте вхождений, контексте или получить помощь и ускоряет процесс работы.

Onboarding

Сайт корпуса стоит на первом месте при поиске в гугле. При открытии страницы справа мы сразу же видим окошко помощи, в котором даны примеры поиска слов и словоформ, а так же синтаксических конструкций и синонимов. Чуть ниже по ссылке доступны более подробные инструкции.

Минималистичный дизайн и удобная организация поиска помогает не запутаться в корпусе. В окошке помощи присутствуют запросы для новичков, но при нажатии на более подробную информацию о поиске по определенным критериям мы можем найти полную инструкцию по использованию ресурса.

Вывод: данный корпус очень прост в использовании новичком, на главной странице дана исчерпывающая информация о запросах и примеры для ознакомления.

Помощь пользователю

По ссылке онлайн-интерфейса корпуса отображается (для незарегистрированных пользователей) заметный блок справочной информации, который можно спрятать. Блок включает в себя ссылки на дополнительные материалы вне онлайн-интерфейса: возможность платной и бесплатной загрузки данных (таких как целые тексты, данные о частоте вхождений тех или иных слов, статистика об n-grams: сочетаниях из n слов, идущих подряд) для офлайн-использования, а также на интерфейс WordAndPhrase, с помощью которого можно найти уже проанализированные данные о частоте вхождений слов в корпусе или обработать целые тексты, которые вводит сам пользователь.

Ниже содержится краткое описание корпуса со ссылками на статистику посещения, сравнение с другими корпусами, разработанными в BYU, данные о балансе корпуса по жанрам, информацию о создании персонализированных виртуальных подкорпусов для работы.
В самом низу блока есть ссылка на таблицу со всеми справочными файлами, включая адаптационный тур с подробным объяснением возможностей поиска, сравнение архитектуры корпусов BYU со SketchEngine и CQP, сравнение COCA с другими крупными

корпусами, работающими с английским языком и с отличиями корпуса от работы с поисковой системой типа Google с целями исследования языка. Также можно получить информацию об использовании корпусов для изучения английского и ссылки на ресурсы BYU с данными о частоте использования слов в языке, collocates (словах, смежных с искомым и частоте вхождений тех или иных сочетаний), n-grams и с полными текстами корпуса.

При переходе между страницами, выполнении поиска окно помощи адаптируется в зависимости от действий и содержит дополнительную информацию о синтаксисе поиска, возможностях сортировки, вывода, сохранения и дальнейшего использования полученных результатов.

Более общую, неспецифичную для COCA информацию о корпусах-проектах BYU можно получить в разделе Help/FAQs, перейдя на главную страницу системы корпусов BYU, где содержатся подробные ответы на 16 вопросов о создании, механизмах работы и идеологии работы корпусов.

Ссылки на справочный материал вынесены на первую страницу, открываемую при работе с корпусом, и более подробные из них легко находятся при переходе с основных блоков информации или по иконкам “информация” или “помощь”, видным с любой страницы, поэтому в корпусе достаточно легко начать ориентироваться, не имея опыта работы с ним. Для начала работы с пониманием базовых механизмов и возможностей достаточно 10-15 минут чтения и практики.

Инструменты

Корпус позволяет задавать сложные запросы с разным количеством параметров.

Можно выбирать, в рамках каких именно жанров, подгрупп или временных отрезках задавать запрос.

Есть возможность создавать и сохранять пользовательские списки связанных каким-либо образом слов и использовать их в поисковых запросах, возможность задавать и сохранять пользовательские подкорпусы по жанрам, датам.

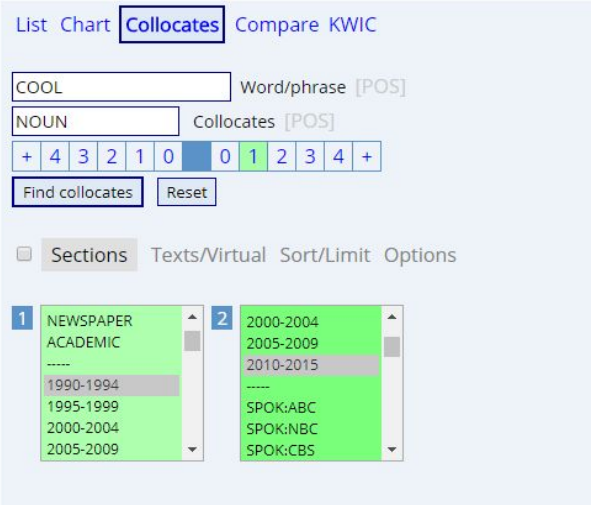
Для результатов в виде списка можно искать фразы определенного типа, конструкции определенного типа с заданным словом/словами, те или иные формы и части речи, возможен поиск по лемме или точной форме, поиск с учетом синонимов, поиск слов некоторого типа, где на заданных местах могут находиться любые буквы.

В зависимости от поиска результаты в контексте или collocates можно сортировать по частоте встречаемости (при этом есть возможность задать минимальную частоту), релевантности (с учетом MI) или алфавиту.

Есть возможность строить сравнительные таблицы для двух и более слов (KWIC, collocates), фраз.

Примеры результатов, которые можно получить, используя эти возможности:

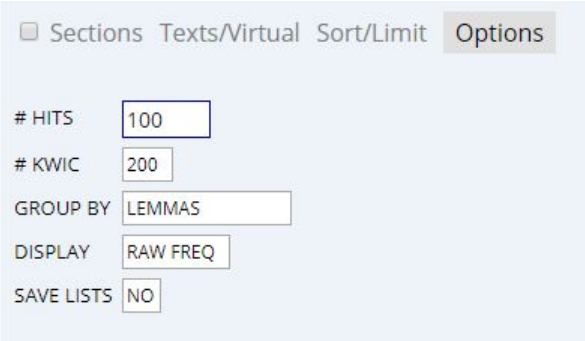
1. Например, нужно узнать, произошли ли какие-то изменения в употреблении слова cool в разные временные периоды. Для этого зададим такой поиск:



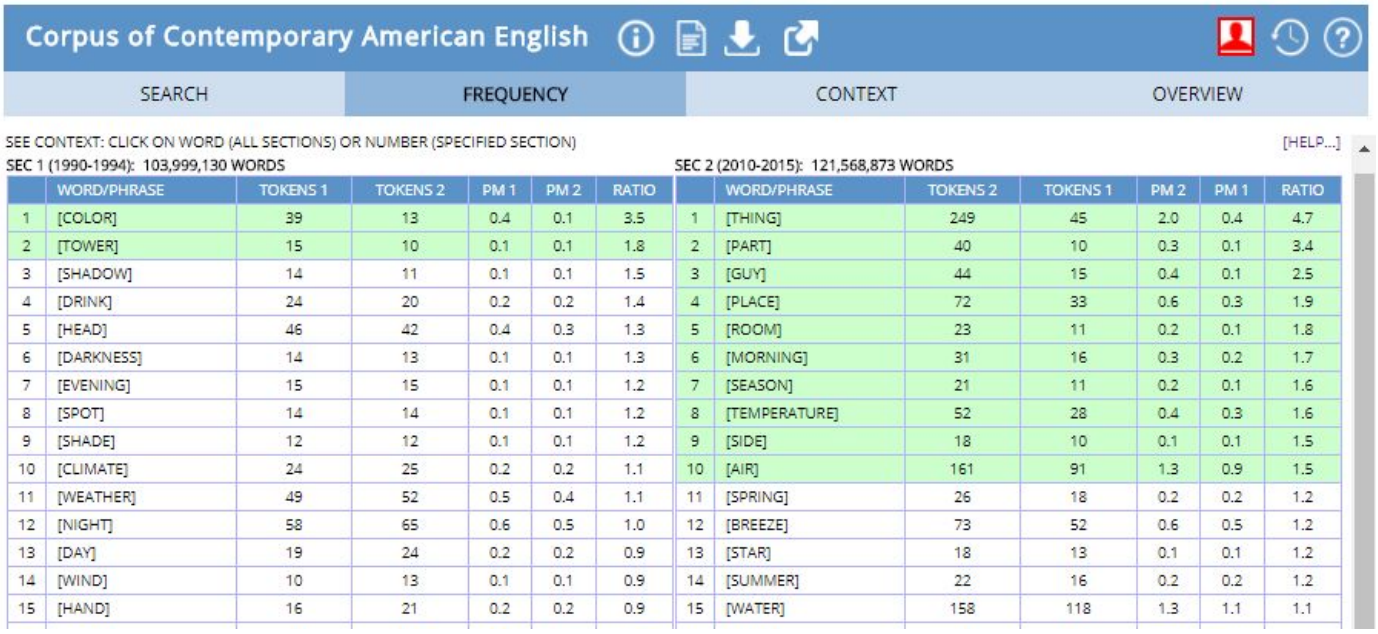
Здеса заданы следующие параметры: collocate должен быть существительным, ближайшим соседом справа, первый диапазон проверки — 1990-1994, второй — 2010-2015.

Параметры сортировки:

Параметры вывода:



На выходе получаем такие результаты:



Corpus of Contemporary American English

SEARCH | FREQUENCY | CONTEXT | OVERVIEW

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

SEC 1 (1990-1994): 103,999,130 WORDS

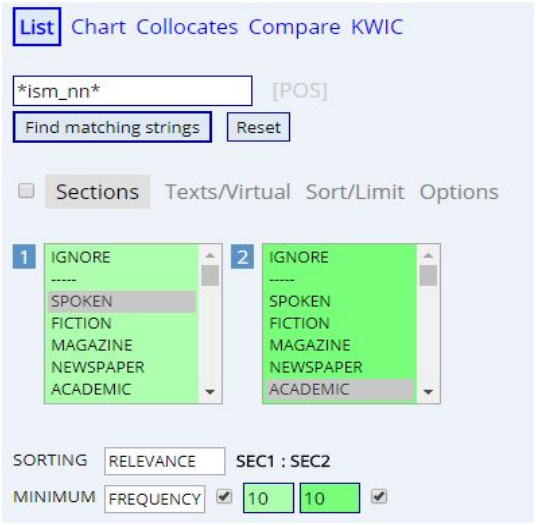
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	[COLOR]	39	13	0.4	0.1	3.5
2	[TOWER]	15	10	0.1	0.1	1.8
3	[SHADOW]	14	11	0.1	0.1	1.5
4	[DRINK]	24	20	0.2	0.2	1.4
5	[HEAD]	46	42	0.4	0.3	1.3
6	[DARKNESS]	14	13	0.1	0.1	1.3
7	[EVENING]	15	15	0.1	0.1	1.2
8	[SPOT]	14	14	0.1	0.1	1.2
9	[SHADE]	12	12	0.1	0.1	1.2
10	[CLIMATE]	24	25	0.2	0.2	1.1
11	[WEATHER]	49	52	0.5	0.4	1.1
12	[NIGHT]	58	65	0.6	0.5	1.0
13	[DAY]	19	24	0.2	0.2	0.9
14	[WIND]	10	13	0.1	0.1	0.9
15	[HAND]	16	21	0.2	0.2	0.9

SEC 2 (2010-2015): 121,568,873 WORDS

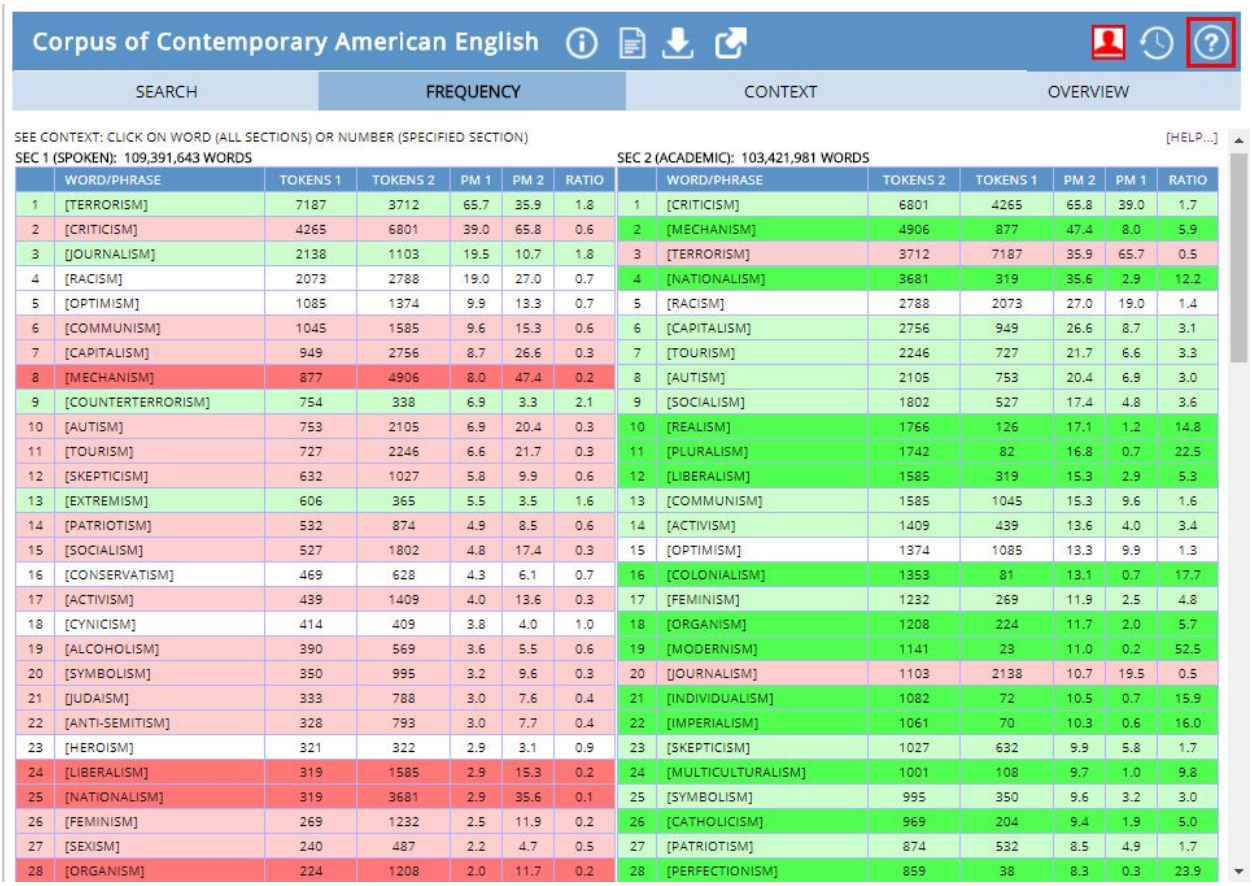
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	[THING]	249	45	2.0	0.4	4.7
2	[PART]	40	10	0.3	0.1	3.4
3	[GUY]	44	15	0.4	0.1	2.5
4	[PLACE]	72	33	0.6	0.3	1.9
5	[ROOM]	23	11	0.2	0.1	1.8
6	[MORNING]	31	16	0.3	0.2	1.7
7	[SEASON]	21	11	0.2	0.1	1.6
8	[TEMPERATURE]	52	28	0.4	0.3	1.6
9	[SIDE]	18	10	0.1	0.1	1.5
10	[AIR]	161	91	1.3	0.9	1.5
11	[SPRING]	26	18	0.2	0.2	1.2
12	[BREEZE]	73	52	0.6	0.5	1.2
13	[STAR]	18	13	0.1	0.1	1.2
14	[SUMMER]	22	16	0.2	0.2	1.2
15	[WATER]	158	118	1.3	1.1	1.1

На основании этого поиска можно сделать предположения об изменении контекста употребления cool от первого периода к последнему (появление наверху списка таких слов как thing, part, guy).

2. Можно пронаблюдать какие существительне на -ism более употребимы в разговорных и академических текстах, задав запрос таким образом:



Результаты поиска (первые 28 строк таблицы):



Corpus of Contemporary American English

SEARCH | FREQUENCY | CONTEXT | OVERVIEW

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

SEC 1 (SPOKEN): 109,391,643 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	[TERRORISM]	7187	3712	65.7	35.9	1.8
2	[CRITICISM]	4265	6801	39.0	65.8	0.6
3	[JOURNALISM]	2138	1103	19.5	10.7	1.8
4	[RACISM]	2073	2788	19.0	27.0	0.7
5	[OPTIMISM]	1085	1374	9.9	13.3	0.7
6	[COMMUNISM]	1045	1585	9.6	15.3	0.6
7	[CAPITALISM]	949	2756	8.7	26.6	0.3
8	[MECHANISM]	877	4906	8.0	47.4	0.2
9	[COUNTERTERRORISM]	754	338	6.9	3.3	2.1
10	[AUTISM]	753	2105	6.9	20.4	0.3
11	[TOURISM]	727	2246	6.6	21.7	0.3
12	[SKEPTICISM]	632	1027	5.8	9.9	0.6
13	[EXTREMISM]	606	365	5.5	3.5	1.6
14	[PATRIOTISM]	532	874	4.9	8.5	0.6
15	[SOCIALISM]	527	1802	4.8	17.4	0.3
16	[CONSERVATISM]	469	628	4.3	6.1	0.7
17	[ACTIVISM]	439	1409	4.0	13.6	0.3
18	[CYNICISM]	414	409	3.8	4.0	1.0
19	[ALCOHOLISM]	390	569	3.6	5.5	0.6
20	[SYMBOLISM]	350	995	3.2	9.6	0.3
21	[JUDAISM]	333	788	3.0	7.6	0.4
22	[ANTI-SEMITISM]	328	793	3.0	7.7	0.4
23	[HEROISM]	321	322	2.9	3.1	0.9
24	[LIBERALISM]	319	1585	2.9	15.3	0.2
25	[NATIONALISM]	319	3681	2.9	35.6	0.1
26	[FEMINISM]	269	1232	2.5	11.9	0.2
27	[SEXISM]	240	487	2.2	4.7	0.5
28	[ORGANISM]	224	1208	2.0	11.7	0.2

SEC 2 (ACADEMIC): 103,421,981 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	[CRITICISM]	6801	4265	65.8	39.0	1.7
2	[MECHANISM]	4906	877	47.4	8.0	5.9
3	[TERRORISM]	3712	7187	35.9	65.7	0.5
4	[NATIONALISM]	3681	319	35.6	2.9	12.2
5	[RACISM]	2788	2073	27.0	19.0	1.4
6	[CAPITALISM]	2756	949	26.6	8.7	3.1
7	[TOURISM]	2246	727	21.7	6.6	3.3
8	[AUTISM]	2105	753	20.4	6.9	3.0
9	[SOCIALISM]	1802	527	17.4	4.8	3.6
10	[REALISM]	1766	126	17.1	1.2	14.8
11	[PLURALISM]	1742	82	16.8	0.7	22.5
12	[LIBERALISM]	1585	319	15.3	2.9	5.3
13	[COMMUNISM]	1585	1045	15.3	9.6	1.6
14	[ACTIVISM]	1409	439	13.6	4.0	3.4
15	[OPTIMISM]	1374	1085	13.3	9.9	1.3
16	[COLONIALISM]	1353	81	13.1	0.7	17.7
17	[FEMINISM]	1232	269	11.9	2.5	4.8
18	[ORGANISM]	1208	224	11.7	2.0	5.7
19	[MODERNISM]	1141	23	11.0	0.2	52.5
20	[JOURNALISM]	1103	2138	10.7	19.5	0.5
21	[INDIVIDUALISM]	1082	72	10.5	0.7	15.9
22	[IMPERIALISM]	1061	70	10.3	0.6	16.0
23	[SKEPTICISM]	1027	632	9.9	5.8	1.7
24	[MULTICULTURALISM]	1001	108	9.7	1.0	9.8
25	[SYMBOLISM]	995	350	9.6	3.2	3.0
26	[CATHOLICISM]	969	204	9.4	1.9	5.0
27	[PATRIOTISM]	874	532	8.5	4.9	1.7
28	[PERFECTIONISM]	859	38	8.3	0.3	23.9

Выводы

COCA является большим, сбалансированным сборником разножанровых текстов американской вариации английского языка.

Он обладает достаточно удобным дизайном и мощным поиском, позволяющим задавать множество параметров и выводить результаты для нескольких слов/периодов/жанров наряду с поиском в контексте или по синонимам. Поиск легко и подробно кастомизируется, можно сохранять списки слов или создавать подкорпуса, внутри которых потом можно вести поиск вместо всего корпуса.

Благодаря этому он очень удобен для сопоставления больших объемов данных — например, сравнения нескольких слов/конструкций по тем или иным признакам, выявления закономерностей каких-то изменений в употреблении слова, нахождения наиболее частотных по употреблению слов, сочетаний или конструкций, подходящих под какие-то параметры.

Кроме того, такой поиск можно использовать при изучении языка: несложно с помощью него выявлять тонкости употребления синонимов в разных контекстах.

Минусами корпуса являются ограничения на объем данных, с которыми можно работать бесплатно, и невозможность работать с результатами поиска в офлайне, как, например, в НКРЯ.