Цель проекта

изучение корпусов *Марка Дэвиса*

с точки зрения практического использования корпусных технологий в деятельности лингвиста

el corpus del español

corpus.byu.edu

corpora, size, queries = better resources, more insight

Задачи проекта

изучить подкорпуса английского и испанского языков;

провести исследование в выбранном подкорпусе по теме дипломной работы с использованием системы запросов

Актуальность проекта

изучение различных аспектов функциони-рования языка в естественном контексте;

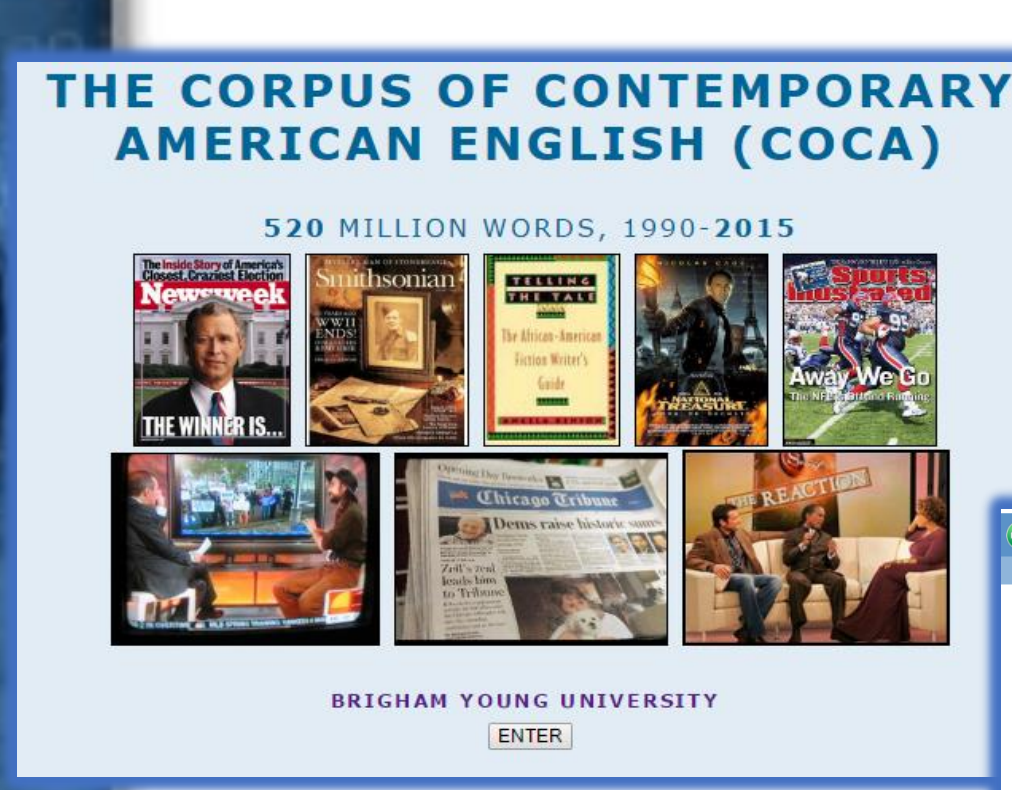
оптимизация процесса исследований

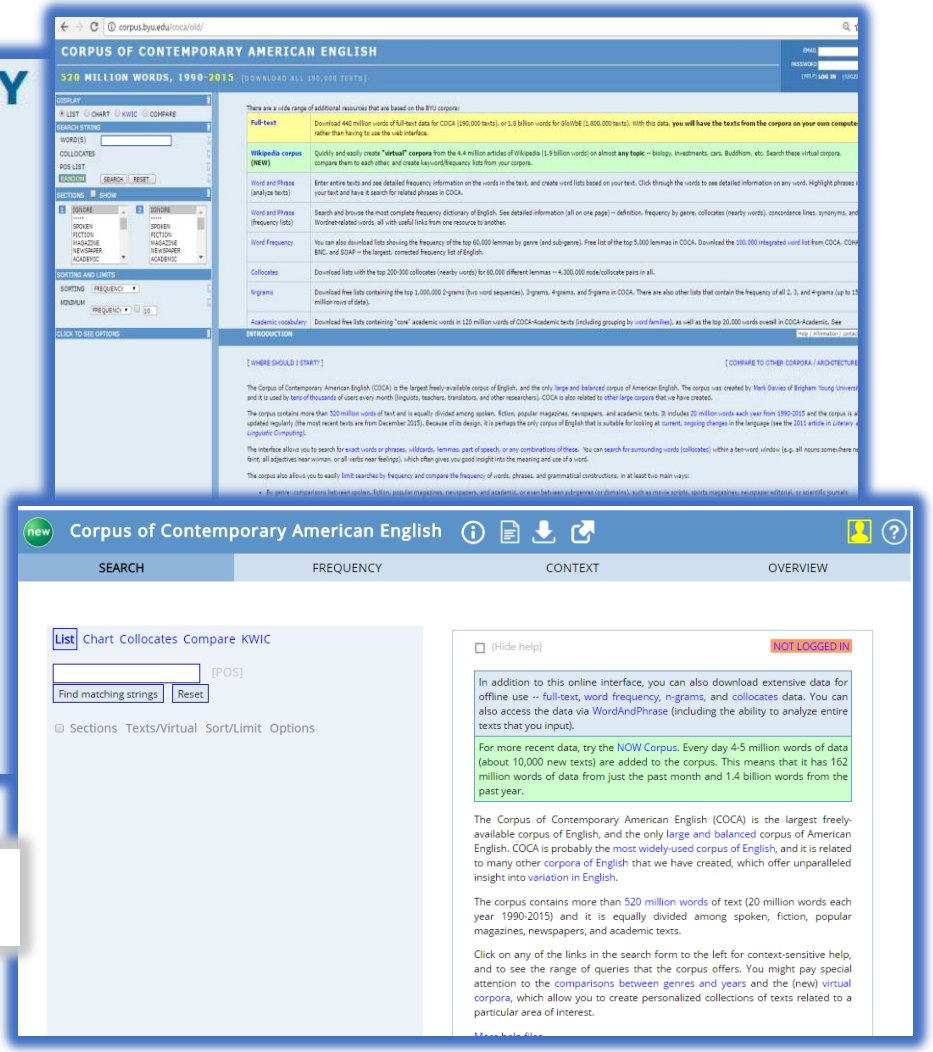
Результаты проекта

- ❑ Описание структуры выбранного подкорпуса
- ❑ Выделение основных характеристик текстов, значимых для дипломного исследования
- ➢ Постановка задачи и проведение исследования с помощью корпуса
- ➢ Подготовка презентации по результатам исследования

BRITISH NATIONAL CORPUS

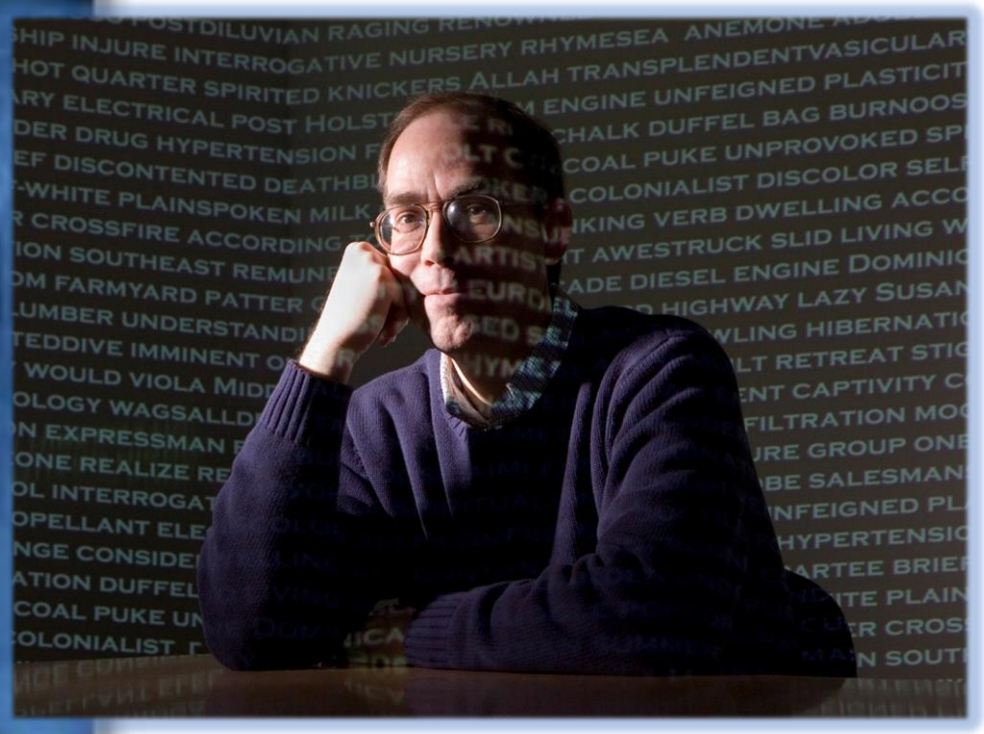
COCA: *Corpus of Contemporary American English*





http://corpus.byu.edu/coca/*old*/

История



Корпус был создан в 2008 году *Марком Дэвисом*, профессором корпусной лингвистики из университета Бригама Янга.

Марк Дэвис собирал и редактировал электронные тексты, спроектировал и внедрил архитектуру корпуса, разработал его веб-интерфейс.

Состав корпуса

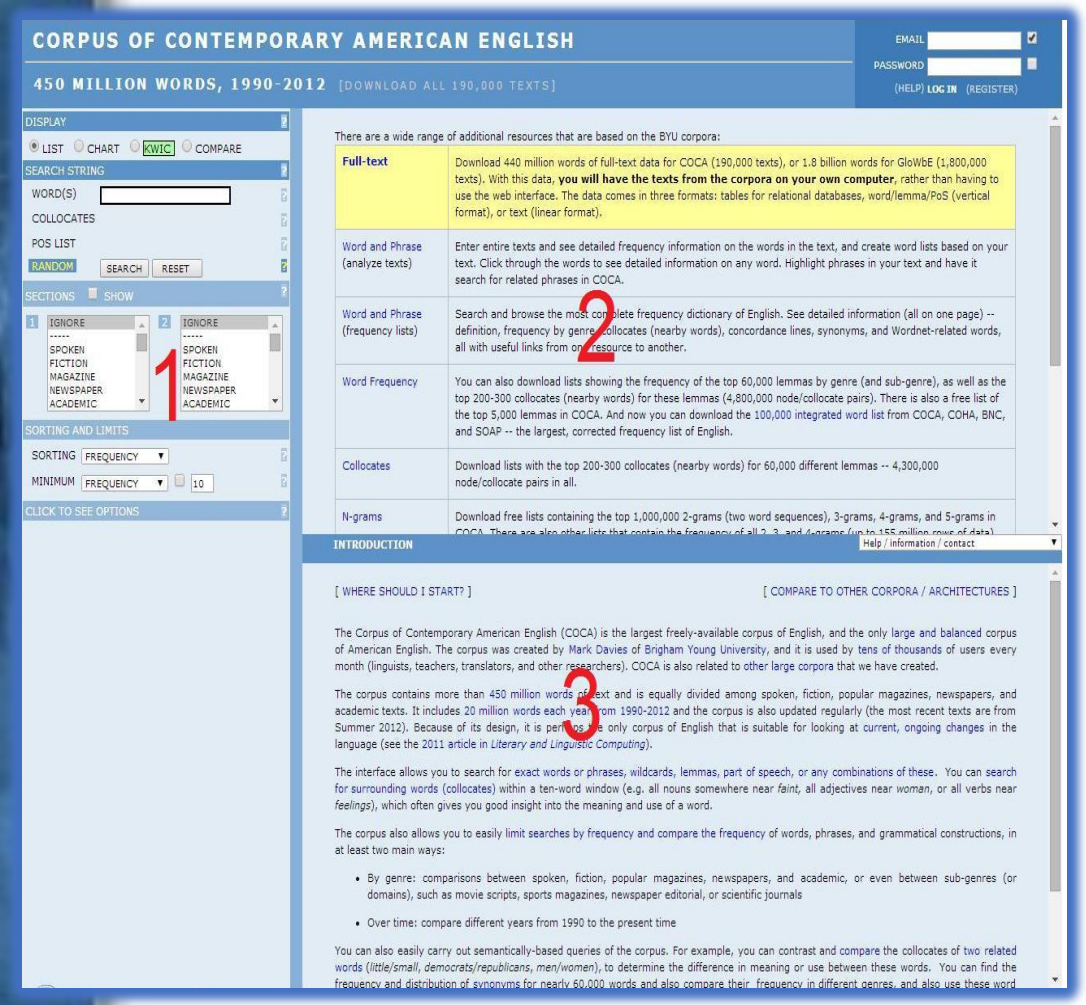
Корпус содержит 520 миллионов слов (190 000 текстов).

Регулярно обновляется (каждый год добавляется порядка 20 миллионов слов), поэтому позволяет исследовать текущие изменения в языке.

В равных долях представлены 5 жанров:

1) ***Устный***: (109 млн. слов). Транскрипция спонтанной речи почти 150 телевизионных программ и радиопередач.
2) ***Художественная литература***: (105 млн. слов) Короткие рассказы и пьесы из литературных, детских и популярных журналов, первые главы первого издания книг с 1990 по настоящее время, сценарии кинофильмов.
3) ***Популярные журналы***: (110 млн. слов). Около 100 журналов различной тематики (новости, здоровье, дом садоводство, женские, финансовые, религиозные и спортивные журналы).
4) ***Газеты***: (106 млн. слов). 10 газет США, тексты взяты из различных разделов (местные новости, оценки, спортивные и финансовые новости).
5) ***Научные журналы***: (103 млн. слов). Почти 100 различных рецензируемых журналов. Охватывают весь диапазон направлений в системе каталога библиотеки Конгресса США, как в целом, так и по количеству слов в год.

Инструкция



Интерфейс корпуса представлен **3** областями:

1) ***область запроса***: вводится запрос, задаются параметры поиска

2) ***списки найденных слов***

3) ***списки конкордансов***

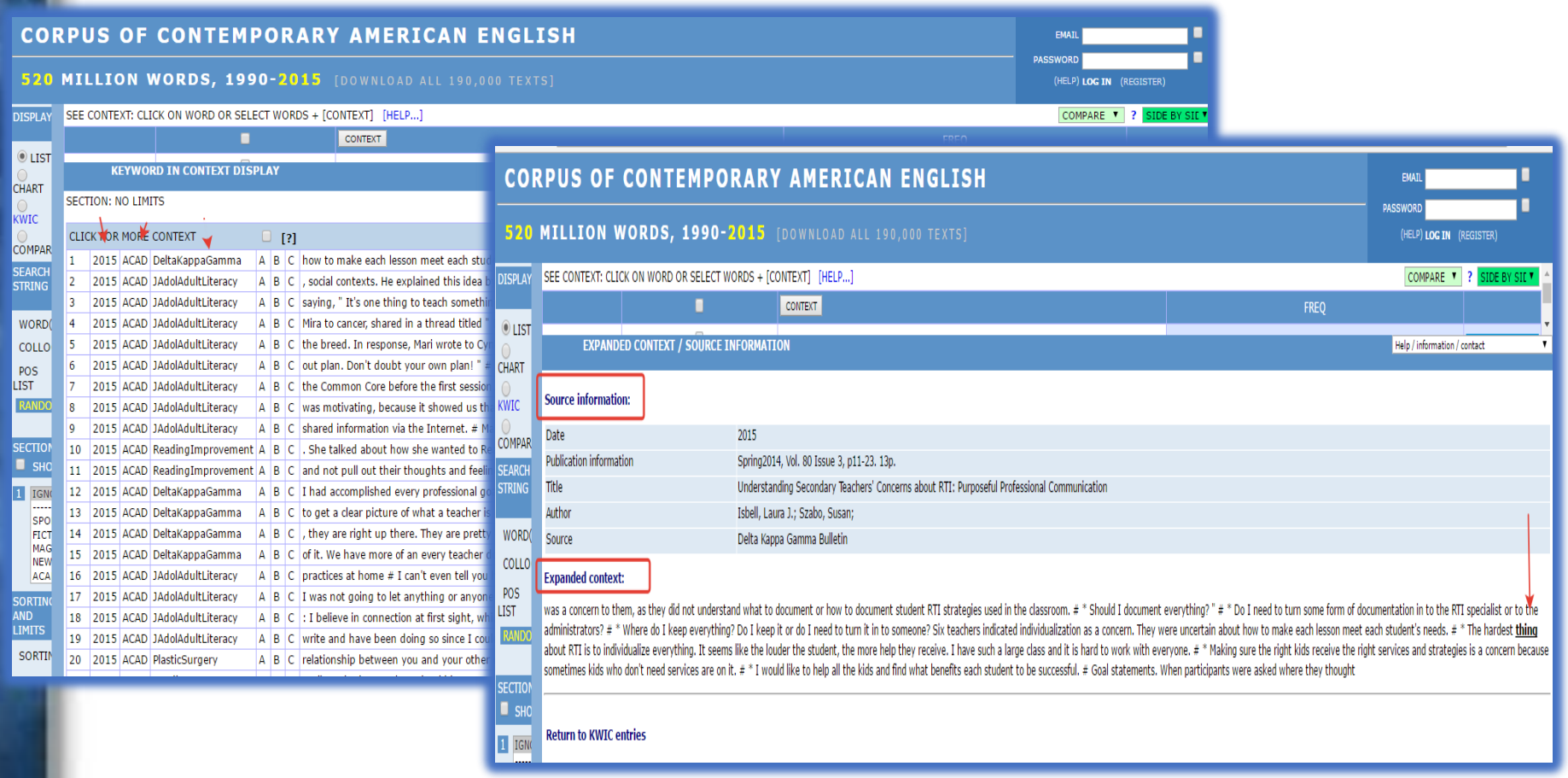
http://corpus.byu.edu/coca/ (*help*)

Простой запрос

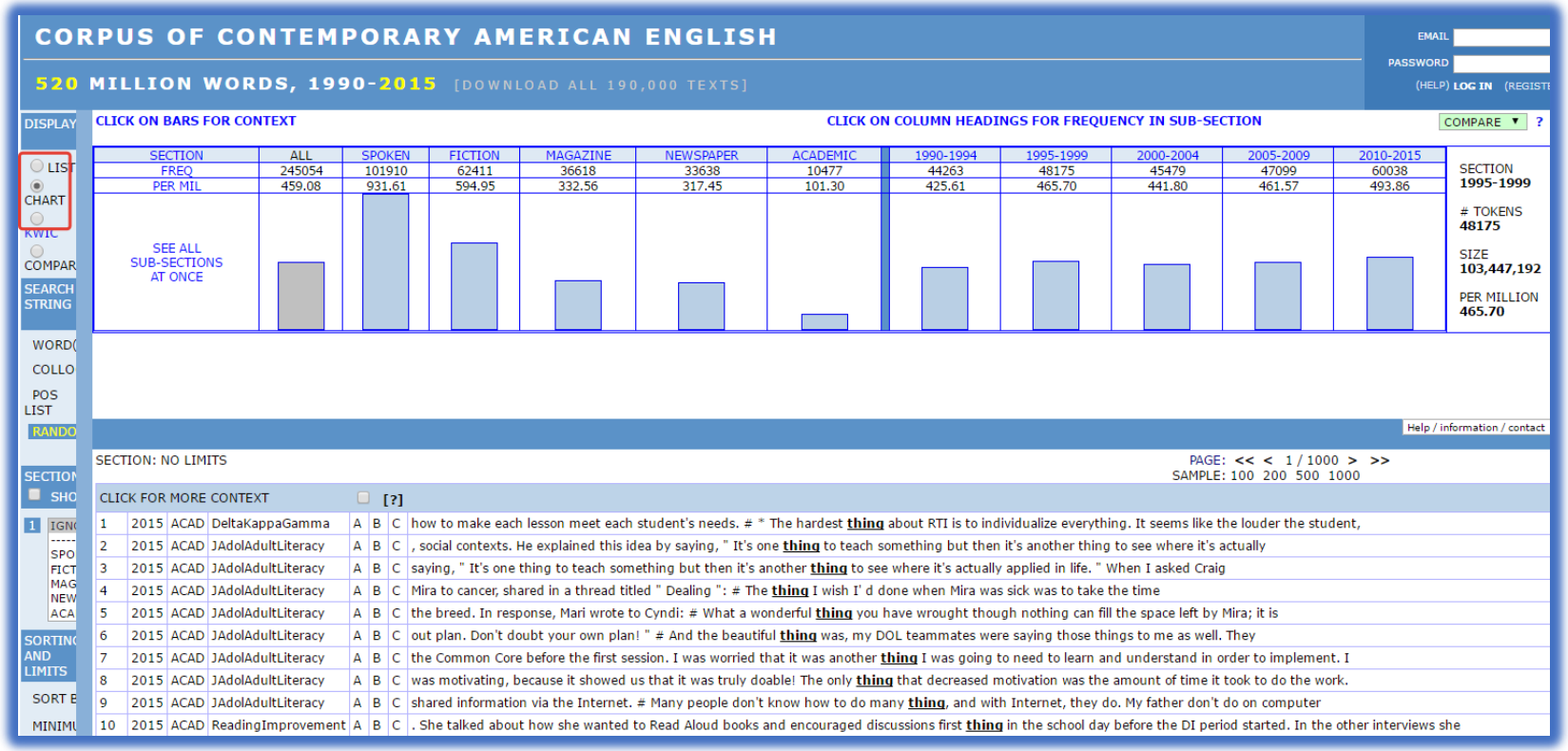
В поле *word(s)* (область запроса) ввести искомое слово – например, ***thing***, и нажать *search*.

В верхней области интерфейса напротив заданного слова отображается цифра, соответствующая общему количеству употреблений этого слова в корпусе (245054).

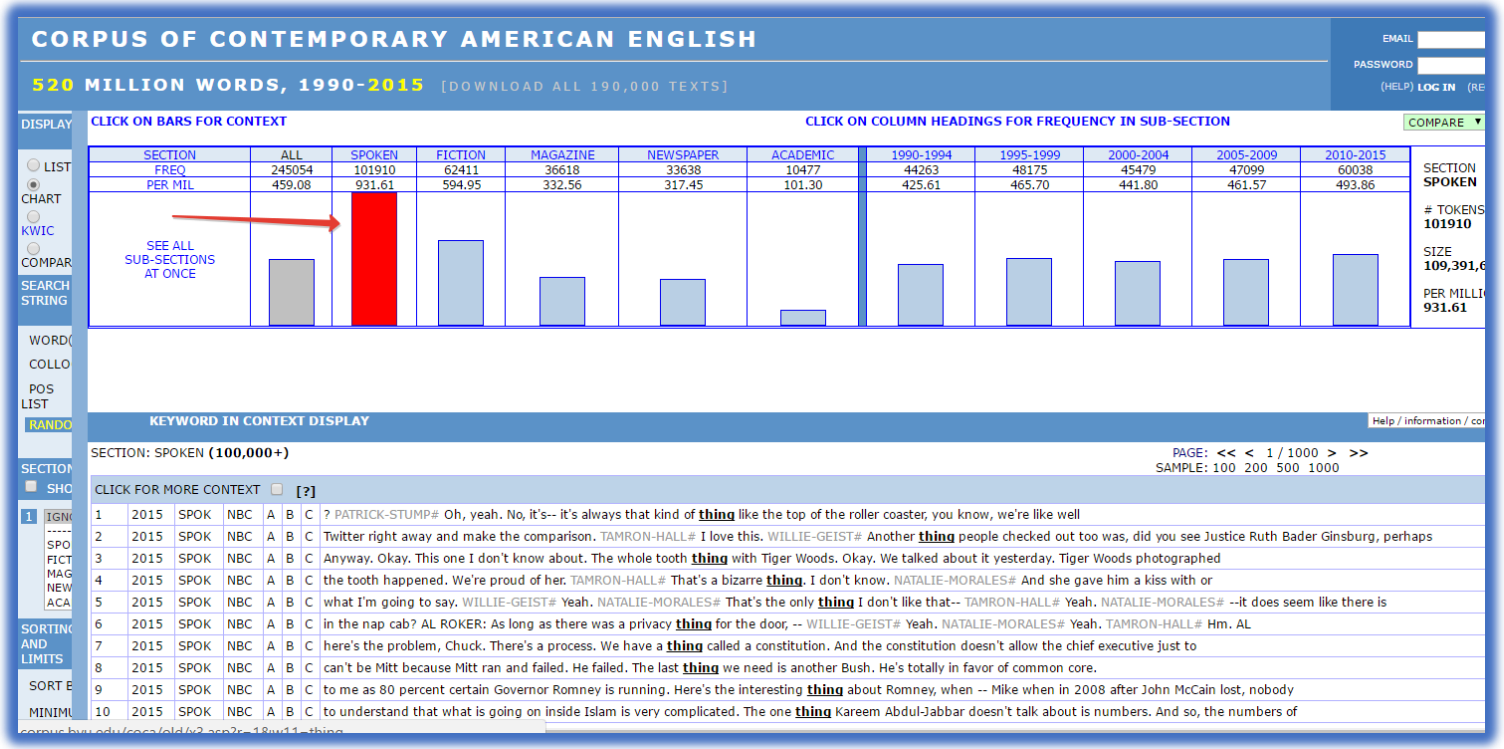
Если кликнуть по заданному слову в области списка найденных слов, то откроется *конкордансный* список. Если кликнуть по любому из первых 4 столбцов, то откроется расширенный контекст и информация о его источнике.



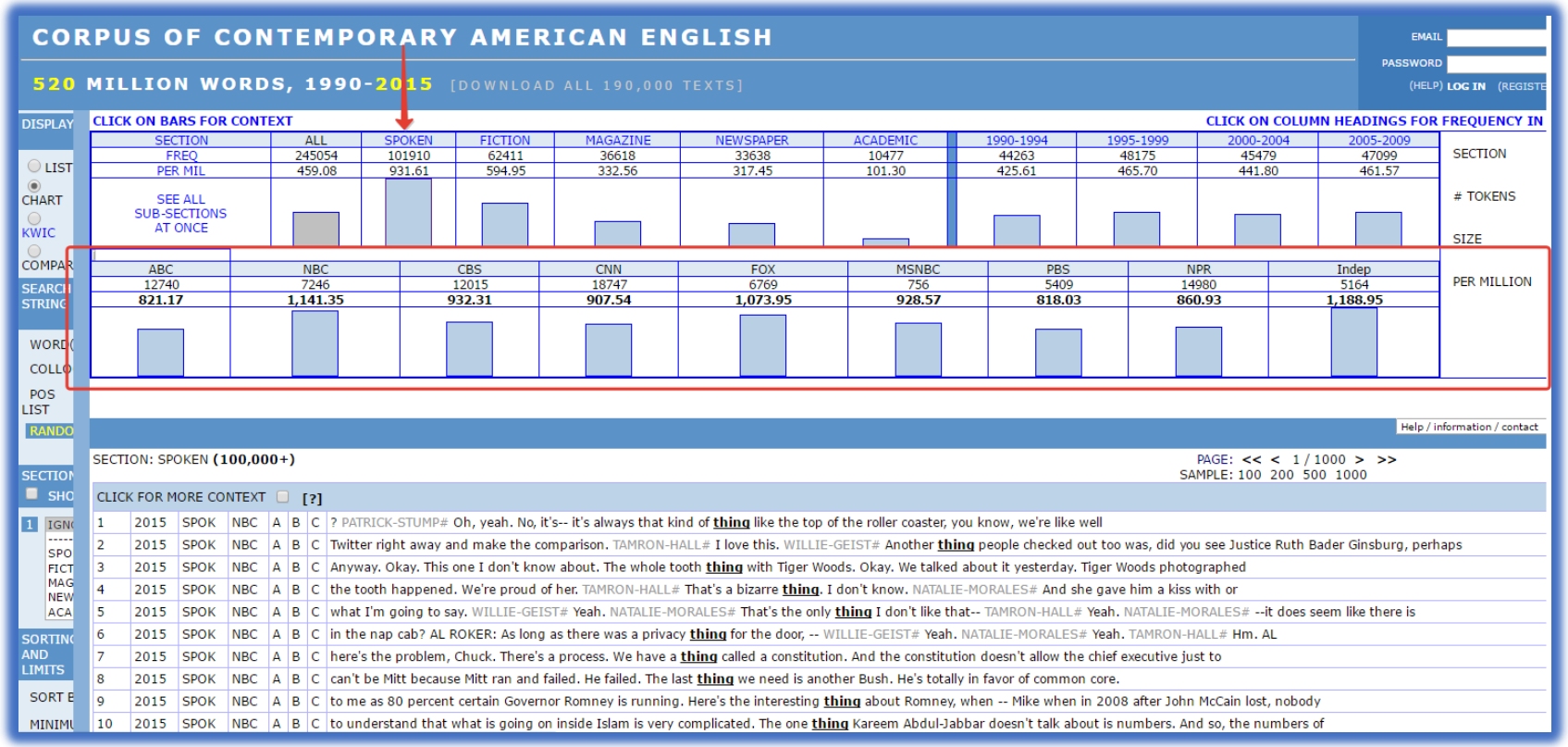
Если в области запроса выбрать *chart* вместо *list*, то будет выдано распределение частоты встречаемости заданного слова в различных жанрах и на различных временных промежутках.



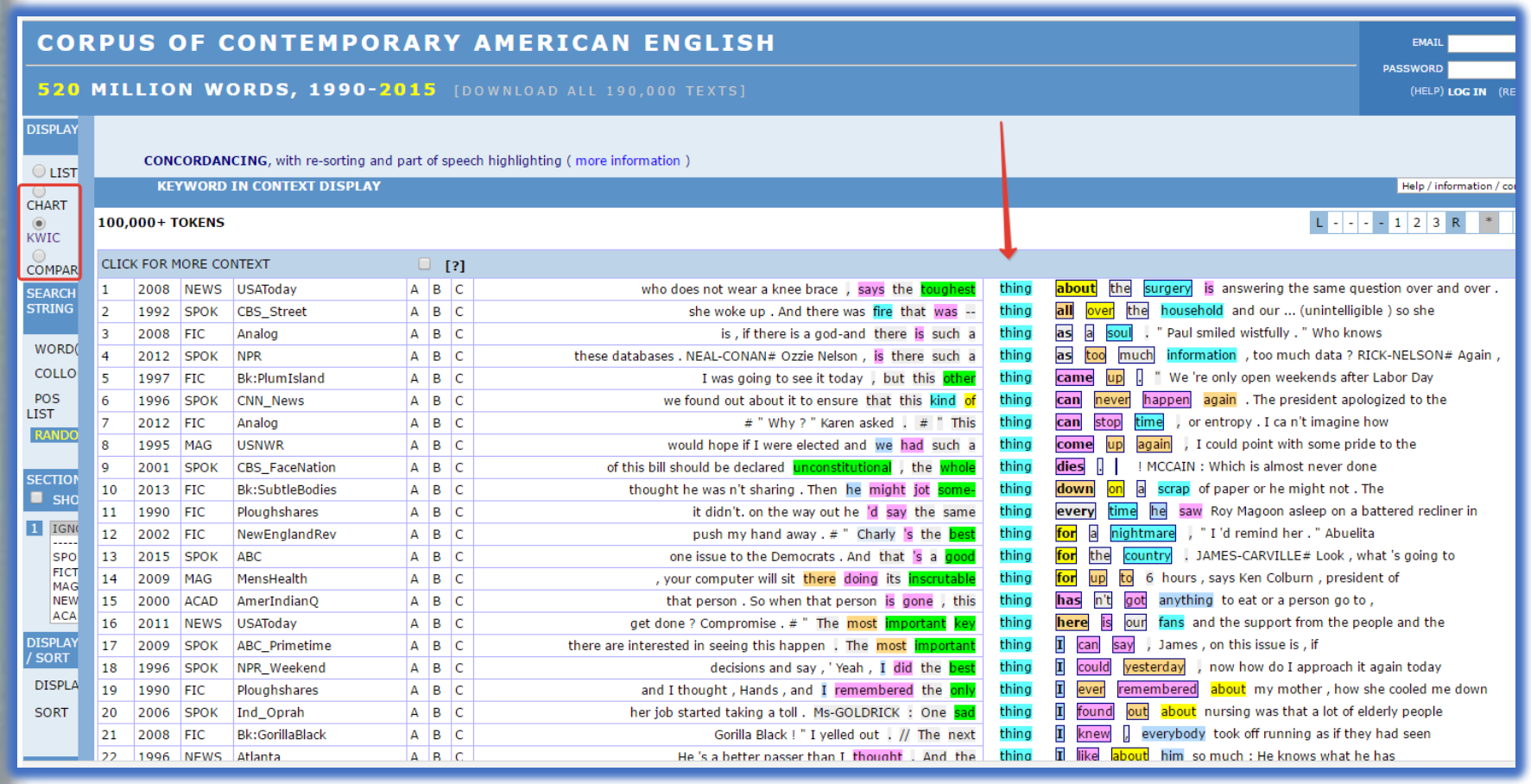
Если кликнуть по ***вертикальному прямоугольнику***, то откроется соответствующий *конкордансный* список.



Если кликнуть на ***название жанра***, то будет выдано распределение *по поджанрам* (например, для устного подкорпуса).



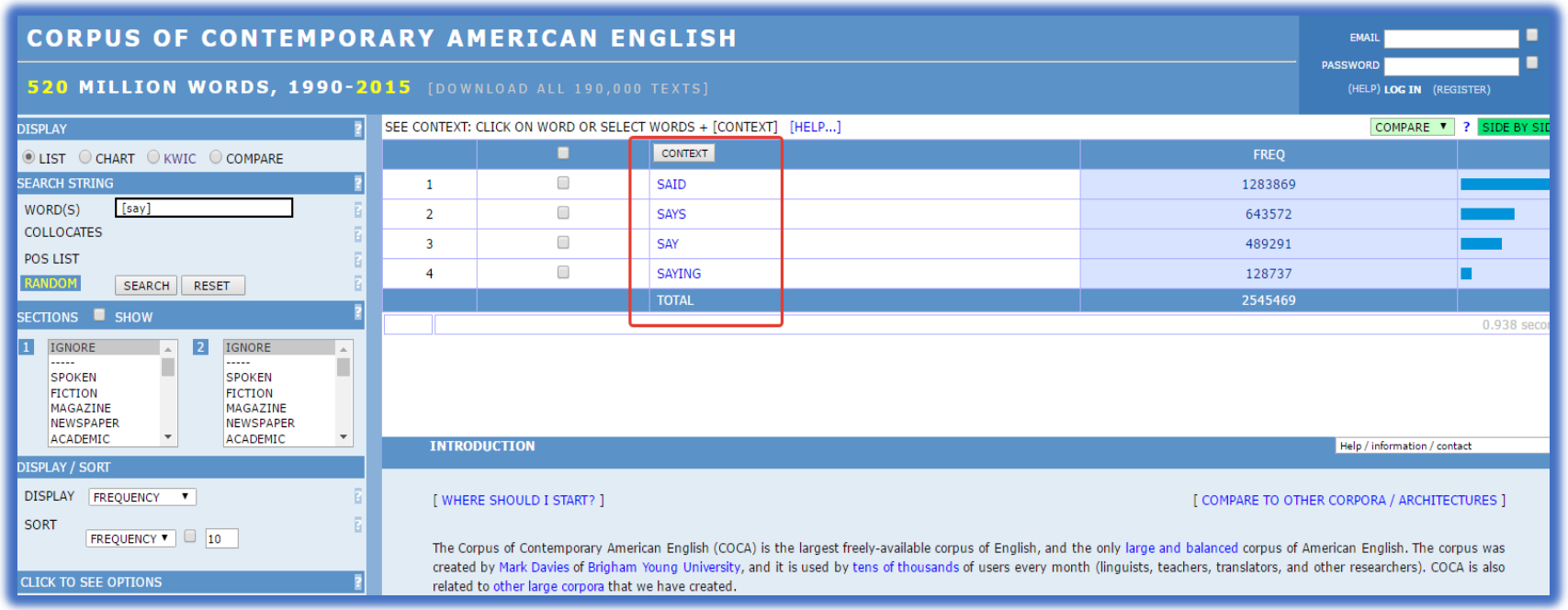
Если в области запроса выбрать «***KWIC***», то будет построен *конкордасный список* по заданному слову (*Key Word In Context*).



Лемматизированный поиск

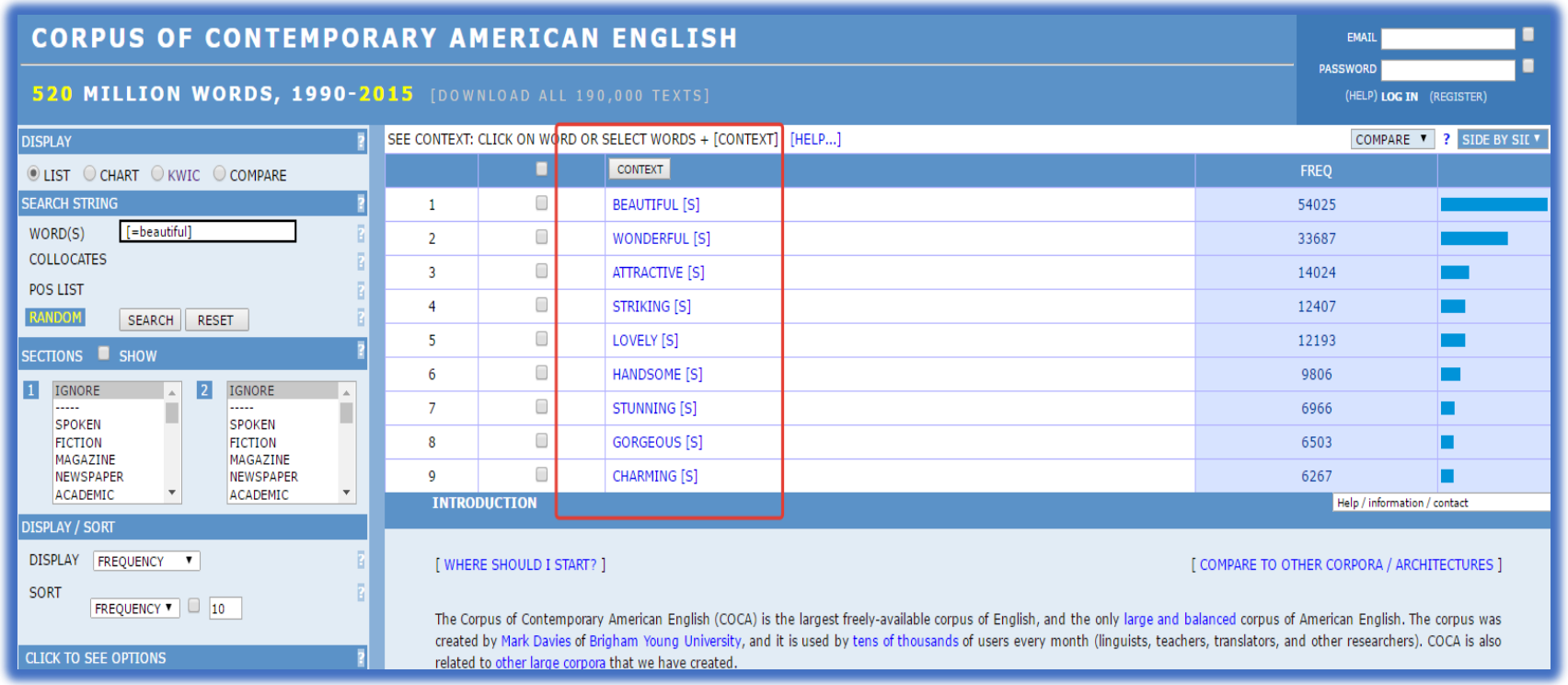
Лемматизированный поиск задается с помощью квадратных скобок [ ].

Например, при введении запроса ***[say]***, в результатах поиска отображаются все его возможные формы – *said*, *says*, *say*, *saying*.



Поиск по синонимам

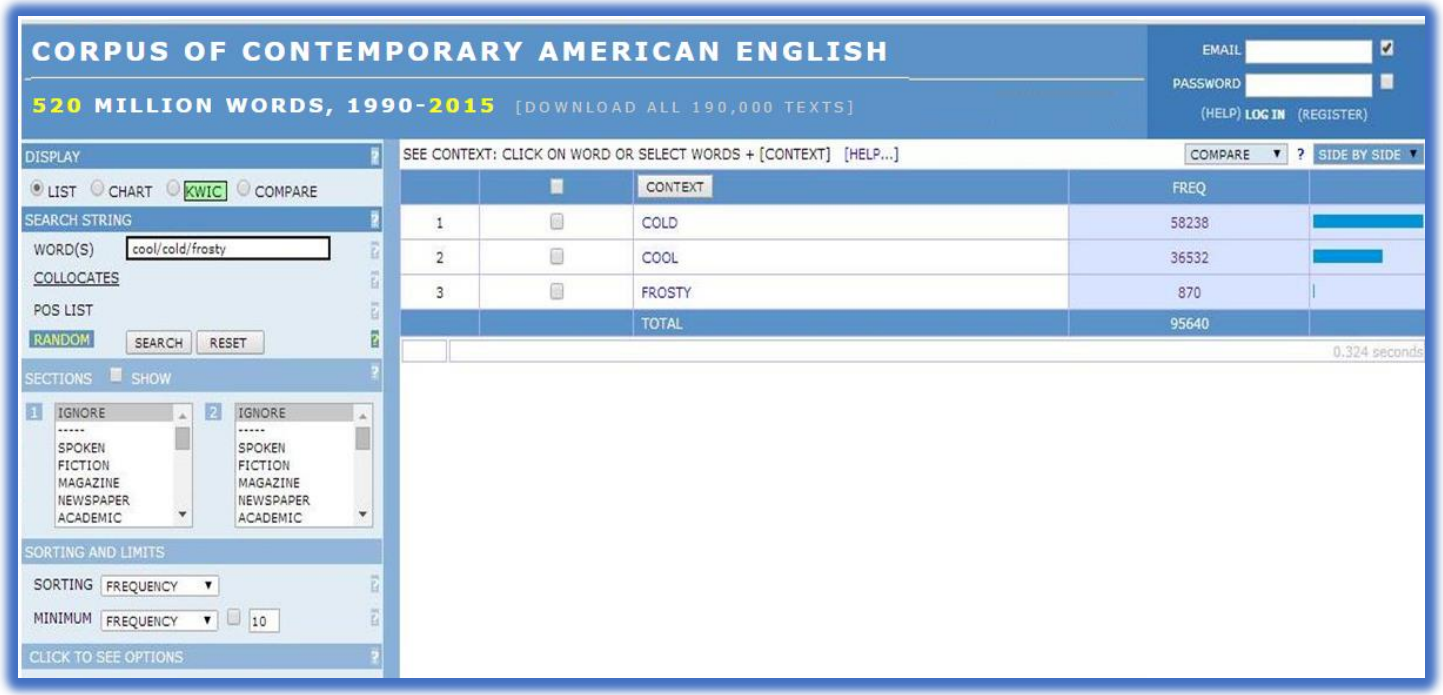
Поиск по синонимам задается следующим образом: [=*слово*]. Например, введя ***[=beautiful]***, в результатах поиска отобразятся его синонимы – *wonderful*, *attractive*, *striking* и т.д.



Поиск по любому из заданных слов

Поиск по любому из заданных слов задается вертикальной (|) либо косой (/) чертой.

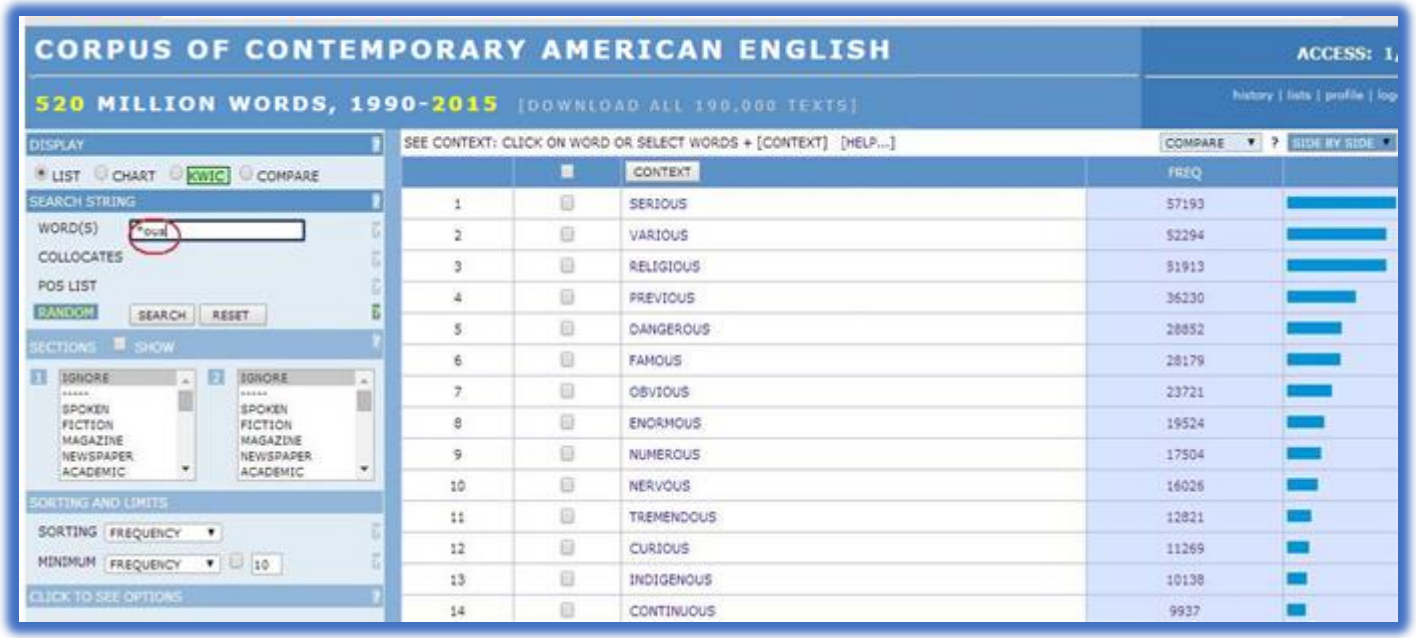
Например: ***cold/cool/frosty***



Подстановочные знаки

В качестве подстановочных знаков используются звездочка (*) - *соответствует любому количеству символов* и
вопросительный знак (**?**) - *соответствует одному символу*.

Например, запрос ***ous**, позволяет выяснить, какие прилагательные с суффиксом «*-ous*» встречаются чаще всего.



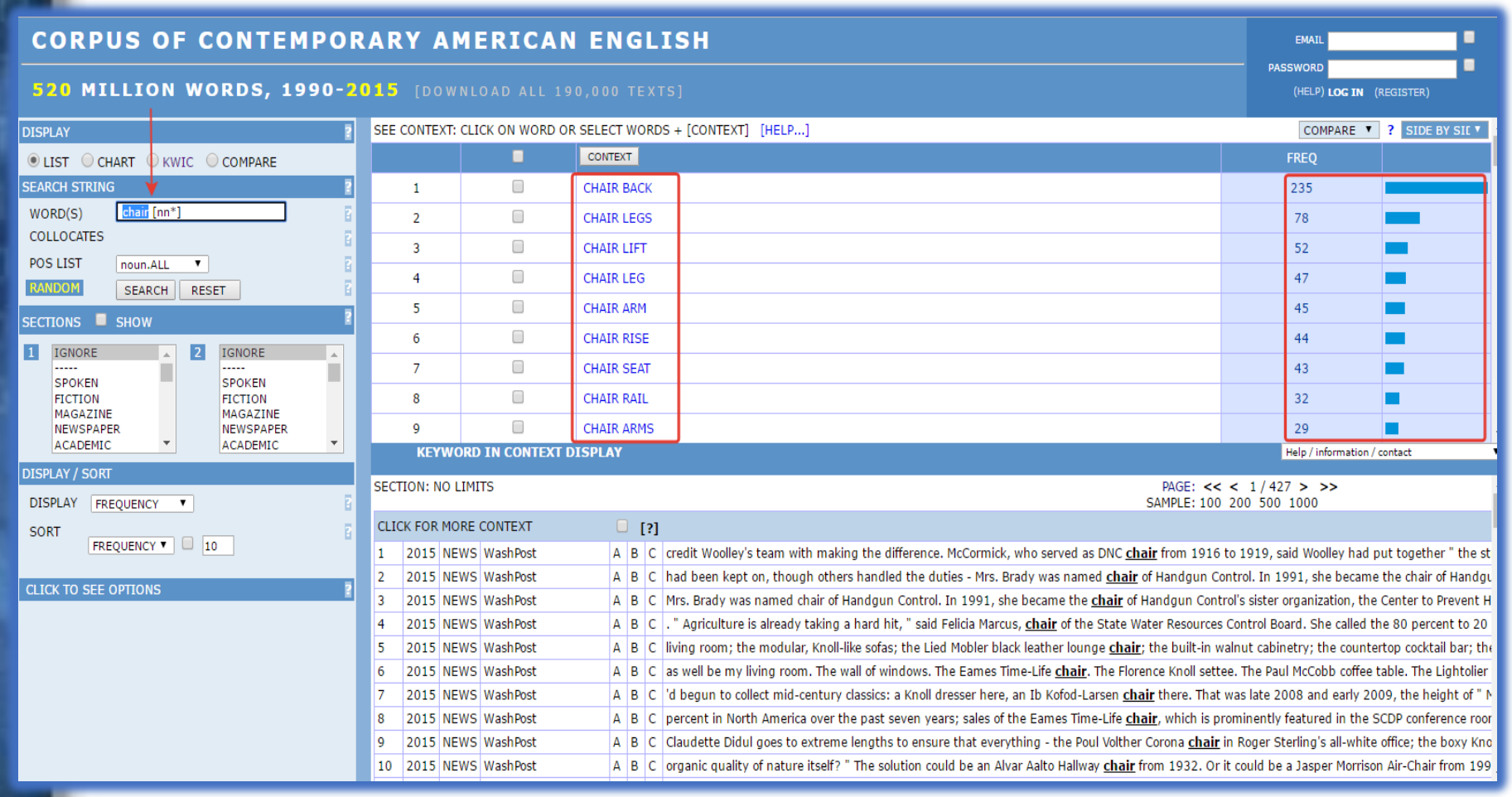
Поиск по частям речи

Для снятия грамматической омонимии необходим *поиск по частям речи*.

Задается следующим образом: нажать на ***POS List***, в выпадающем списке выбрать необходимую часть речи - соответствующий тег автоматически появится в строке запроса через пробел после заданного слова.

Пробел необходимо заменить на *точку* (.)

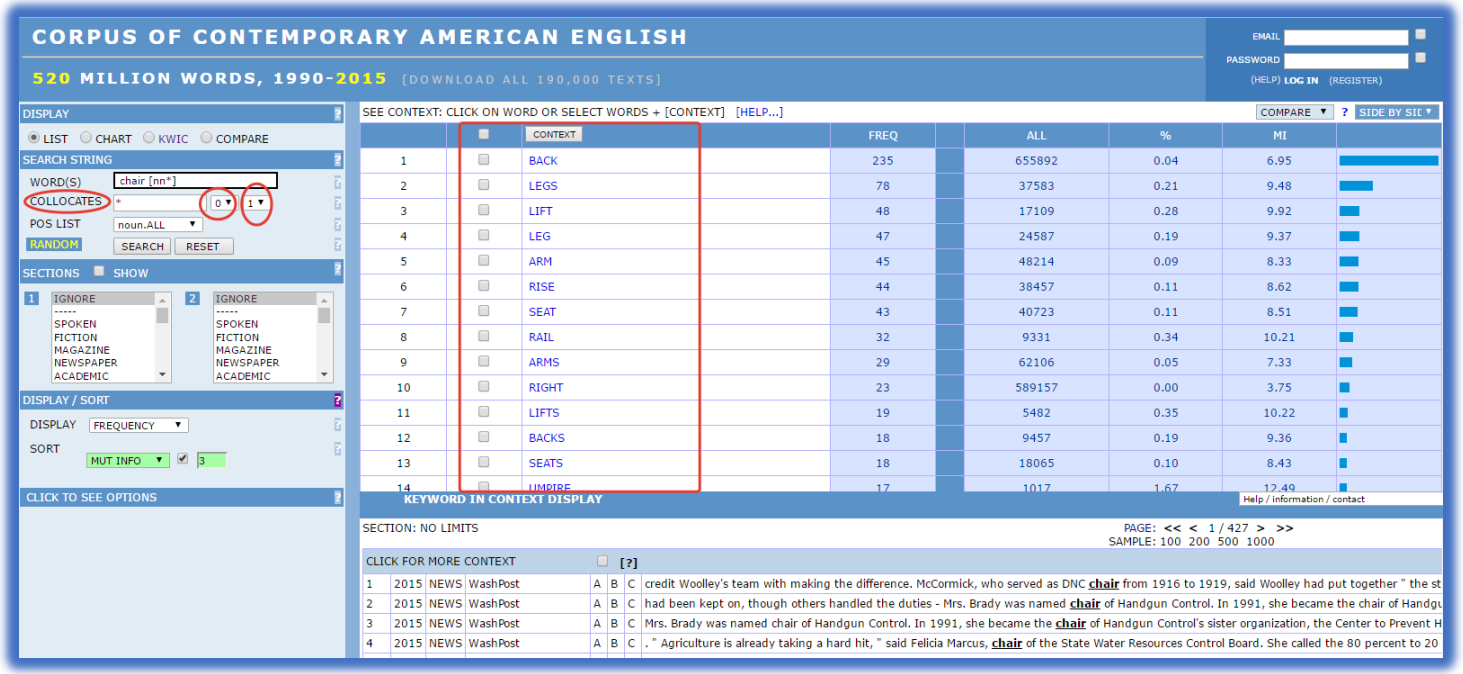
Если пробел оставить (не менять на точку), то будут выведены существительные, которые чаще всего следуют за заданным словом.



Поиск по коллокациям

Этот же результат будет выведен, если воспользоваться поиском по коллокациям (соседним словам).

Необходимо нажать ***collocates***, в выпадающем списке ***POS List*** выбрать тег нужной части речи, задать интервал - ***0*** слов слева и ***1*** слово справа от заданного слова.

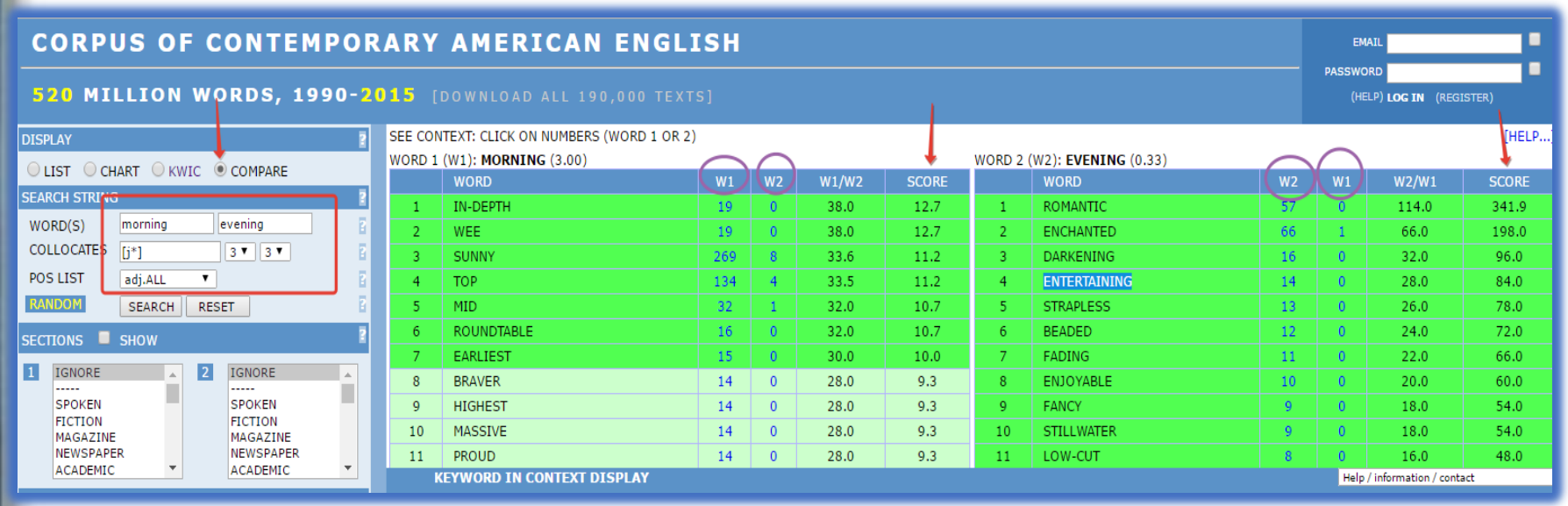


Сравнительный поиск

Например, сравним, какие прилагательные чаще всего сопутствуют слову ***evening***, а какие - слову ***morning***.

Числовые значения столбцов ***W1***, ***W2*** обозначают *общее количество* вхождений для каждого прилагательного, сопутствующего заданным словам.

Слова сортированы по релевантности, которая определяется количеством взаимной информации (столбец ***score***).



	WORD	W1	W2	W1/W2	SCORE		WORD	W2	W1	W2/W1	SCORE
1	IN-DEPTH	19	0	38.0	12.7	1	ROMANTIC	57	0	114.0	341.9
2	WEE	19	0	38.0	12.7	2	ENCHANTED	66	1	66.0	198.0
3	SUNNY	269	8	33.6	11.2	3	DARKENING	16	0	32.0	96.0
4	TOP	134	4	33.5	11.2	4	ENTERTAINING	14	0	28.0	84.0
5	MID	32	1	32.0	10.7	5	STRAPLESS	13	0	26.0	78.0
6	ROUNDTABLE	16	0	32.0	10.7	6	BEADED	12	0	24.0	72.0
7	EARLIEST	15	0	30.0	10.0	7	FADING	11	0	22.0	66.0
8	BRAVER	14	0	28.0	9.3	8	ENJOYABLE	10	0	20.0	60.0
9	HIGHEST	14	0	28.0	9.3	9	FANCY	9	0	18.0	54.0
10	MASSIVE	14	0	28.0	9.3	10	STILLWATER	9	0	18.0	54.0
11	PROUD	14	0	28.0	9.3	11	LOW-CUT	8	0	16.0	48.0