



Цель проекта

BYU-BNC
BRITISH NATIONAL CORPUS
100 MILLION WORDS, UK, 1980s-1993

BRITISH NATIONAL CORPUS

BRIGHAM YOUNG UNIVERSITY

THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA)
520 MILLION WORDS, 1990-2015



BRIGHAM YOUNG UNIVERSITY

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

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**

COCA: Corpus of Contemporary American English

THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA)

520 MILLION WORDS, 1990-2015



BRIGHAM YOUNG UNIVERSITY

ENTER

<http://corpus.byu.edu/coca/old/>

There are a wide range of additional resources that are based on the BYU corpus. You can search for words, phrases, and collocations. You can also search for words and phrases in the corpus. You can also search for words and phrases in the corpus. You can also search for words and phrases in the corpus.

- Full-text**: Download 480 million words of full-text data for COCA (300,000 texts), or 1.8 billion words for NOW (1,800,000 texts), with this data, you will have the texts from the corpora on your own computer rather than having to use the web interface.
- Virtual corpora (NEW)**: Quickly and easily create "virtual" corpora from the 4.4 million articles of Wikipedia (3.9 billion words) on almost any topic – biology, investments, cars, Buddhism, etc. Search these virtual corpora, compare them to each other, and create key-word frequency lists from your corpora.
- Word and Phrase (Spaice texts)**: Enter entire texts and see detailed frequency information on the words in the text, and create word lists based on your text. Click through the words to see detailed information on any word, highlight phrases in your text and have it search for related phrases in COCA.
- Word and Phrase (Frequency lists)**: Search and browse the most complete frequency dictionary of English. See detailed information (all on one page) – definition, frequency by genre, collocations (nearby words), concordance lines, synonyms, and word-related words, all with useful links from one resource to another.
- Word Frequency**: You can also download lists showing the frequency of the top 60,000 lemmas by genre (and subgenre). Free list of the top 5,000 lemmas in COCA. Download the 300,000 integrated word list from COCA, COCA-BNC, and SOEP – the largest, corrected frequency list of English.
- Collocates**: Download lists with the top 200-300 collocates (nearby words) for 60,000 different lemmas – 4,300,000 node/collocate pairs in all.
- n-grams**: Download free lists containing the top 1,000,000 2-grams (two-word sequences), 3-grams, 4-grams, and 5-grams in COCA. There are also other lists that contain the frequency of all 2-, 3-, and 4-grams (up to 10 million rows of data).
- Academic vocabulary**: Download free lists containing "low" academic words in 120 million words of COCA-academic texts (including grouping by word families), as well as the top 20,000 words overall in COCA-Academic. See [http://www.coca.academic.org](#)

new Corpus of Contemporary American English

SEARCH FREQUENCY CONTEXT OVERVIEW

List Chart Collocates Compare KWIC

[] (Hide help) NOT LOGGED IN

In addition to this online interface, you can also download extensive data for offline use -- full-text, word frequency, n-grams, and collocates data. You can also access the data via WordAndPhrase (including the ability to analyze entire texts that you input).

For more recent data, try the NOW Corpus. Every day 4-5 million words of data (about 10,000 new texts) are added to the corpus. This means that it has 162 million words of data from just the past month and 1.4 billion words from the past year.

The Corpus of Contemporary American English (COCA) is the largest freely available corpus of English, and the only large and balanced corpus of American English. COCA is probably the most widely used corpus of English, and it is related to many other corpora of English that we have created, which offer unparalleled insight into variation in English.

The corpus contains more than 520 million words of text (20 million words each year 1990-2015) and it is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts.

Click on any of the links in the search form to the left for context-sensitive help, and to see the range of queries that the corpus offers. You might pay special attention to the comparisons between genres and years and the (new) virtual corpora, which allow you to create personalized collections of texts related to a particular area of interest.



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

CORPUS OF CONTEMPORARY AMERICAN ENGLISH
450 MILLION WORDS, 1990-2012 [DOWNLOAD ALL 190,000 TEXTS]

EMAIL
PASSWORD
(HELP) LOG IN (REGISTER)

DISPLAY
LIST CHART KWIC COMPARE

SEARCH STRING
WORD(S)
COLLOCATES
POS LIST
RANDOM SEARCH RESET

SECTIONS SHOW
1 IGNORE 2 IGNORE
SPOKEN SPOKEN
FICTION FICTION
MAGAZINE MAGAZINE
NEWSPAPER NEWSPAPER
ACADEMIC ACADEMIC

SORTING AND LIMITS
SORTING FREQUENCY
MINIMUM FREQUENCY 10
CLICK TO SEE OPTIONS

There are a wide range of additional resources that are based on the BYU corpora:

Full-text	Download 440 million words of full-text data for COCA (190,000 texts), or 1.8 billion words for GloWbE (1,800,000 texts). With this data, you will have the texts from the corpora on your own computer , rather than having to use the web interface. The data comes in three formats: tables for relational databases, word/lemma/PoS (vertical format), or text (linear format).
Word and Phrase (analyze texts)	Enter entire texts and see detailed frequency information on the words in the text, and create word lists based on your text. Click through the words to see detailed information on any word. Highlight phrases in your text and have it search for related phrases in COCA.
Word and Phrase (frequency lists)	Search and browse the most complete frequency dictionary of English. See detailed information (all on one page) -- definition, frequency by genre, collocates (nearby words), concordance lines, synonyms, and Wordnet-related words, all with useful links from one resource to another.
Word Frequency	You can also download lists showing the frequency of the top 60,000 lemmas by genre (and sub-genre), as well as the top 200-300 collocates (nearby words) for these lemmas (4,800,000 node/collocate pairs). There is also a free list of the top 5,000 lemmas in COCA. And now you can download the 100,000 integrated word list from COCA, COHA, BNC, and SOAP -- the largest, corrected frequency list of English.
Collocates	Download lists with the top 200-300 collocates (nearby words) for 60,000 different lemmas -- 4,300,000 node/collocate pairs in all.
N-grams	Download free lists containing the top 1,000,000 2-grams (two word sequences), 3-grams, 4-grams, and 5-grams in COCA. There are also other lists that contain the frequency of all 2-, 3-, and 4-grams (up to 185 million rows of data).

INTRODUCTION
[WHERE SHOULD I START?] [COMPARE TO OTHER CORPORA / ARCHITECTURES]

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, and the only large and balanced corpus of American English. The corpus was created by Mark Davies of Brigham Young University, and it is used by tens of thousands of users every month (linguists, teachers, translators, and other researchers). COCA is also related to other large corpora that we have created.

The corpus contains more than 450 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. It includes 20 million words each year from 1990-2012 and the corpus is also updated regularly (the most recent texts are from Summer 2012). Because of its design, it is perhaps the only corpus of English that is suitable for looking at current, ongoing changes in the language (see the 2011 article in *Literary and Linguistic Computing*).

The interface allows you to search for exact words or phrases, wildcards, lemmas, part of speech, or any combinations of these. You can search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near faint, all adjectives near woman, or all verbs near feelings), which often gives you good insight into the meaning and use of a word.

The corpus also allows you to easily limit searches by frequency and compare the frequency of words, phrases, and grammatical constructions, in at least two main ways:

- By genre: comparisons between spoken, fiction, popular magazines, newspapers, and academic, or even between sub-genres (or domains), such as movie scripts, sports magazines, newspaper editorial, or scientific journals
- Over time: compare different years from 1990 to the present time

You can also easily carry out semantically-based queries of the corpus. For example, you can contrast and compare the collocates of two related words (*little/small, democrats/republicans, men/women*), to determine the difference in meaning or use between these words. You can find the frequency and distribution of synonyms for nearly 60,000 words and also compare their frequency in different genres, and also use these word

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).

The screenshot shows the COCA search interface. The search term "thing" is entered in the "WORD(S)" field. The search results table shows "THING" with a frequency of 245054. The interface includes various search options like "LIST", "CHART", "KWIC", and "COMPARE", and a sidebar with "SECTIONS" and "SORTING AND LIMITS".

	CONTEXT	FREQ
1	THING	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.

The image shows two overlapping screenshots of the COCA website. The background screenshot shows a search result for the word "thing" in a KWIC list. The foreground screenshot shows the expanded context for the word "thing" in the 12th row of the KWIC list, including source information and the full sentence.

COCA Website Interface:

- Header: CORPUS OF CONTEMPORARY AMERICAN ENGLISH
- Search Results: 520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]
- Navigation: DISPLAY, SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...], COMPARE, SIDE BY SIDE
- Left Panel: LIST, CHART, KWIC, COMPARE, SEARCH STRING, WORD, COLLO, POS LIST, RANDO, SECTION, SHC, SORTING AND LIMITS, SORTING

KWIC List (Background Screenshot):

1	2015	ACAD	DeltaKappaGamma	A	B	C	how to make each lesson meet each stu		
2	2015	ACAD	JAdolAdultLiteracy	A	B	C	, social contexts. He explained this idea		
3	2015	ACAD	JAdolAdultLiteracy	A	B	C	saying, " It's one thing to teach someth		
4	2015	ACAD	JAdolAdultLiteracy	A	B	C	Mira to cancer, shared in a thread titled		
5	2015	ACAD	JAdolAdultLiteracy	A	B	C	the breed. In response, Mari wrote to C		
6	2015	ACAD	JAdolAdultLiteracy	A	B	C	out plan. Don't doubt your own plan!"		
7	2015	ACAD	JAdolAdultLiteracy	A	B	C	the Common Core before the first sessio		
8	2015	ACAD	JAdolAdultLiteracy	A	B	C	was motivating, because it showed us th		
9	2015	ACAD	JAdolAdultLiteracy	A	B	C	shared information via the Internet. # M		
10	2015	ACAD	ReadingImprovement	A	B	C	. She talked about how she wanted to R		
11	2015	ACAD	ReadingImprovement	A	B	C	and not pull out their thoughts and feel		
12	2015	ACAD	DeltaKappaGamma	A	B	C	I had accomplished every professional g		
13	2015	ACAD	DeltaKappaGamma	A	B	C	to get a clear picture of what a teacher i		
14	2015	ACAD	DeltaKappaGamma	A	B	C	, they are right up there. They are prett		
15	2015	ACAD	DeltaKappaGamma	A	B	C	of it. We have more of an every teacher		
16	2015	ACAD	JAdolAdultLiteracy	A	B	C	practices at home # I can't even tell you		
17	2015	ACAD	JAdolAdultLiteracy	A	B	C	I was not going to let anything or anyon		
18	2015	ACAD	JAdolAdultLiteracy	A	B	C	: I believe in connection at first sight, w		
19	2015	ACAD	JAdolAdultLiteracy	A	B	C	write and have been doing so since I co		
20	2015	ACAD	PlasticSurgery	A	B	C	relationship between you and your other		

Expanded Context / Source Information (Foreground Screenshot):

Source information:

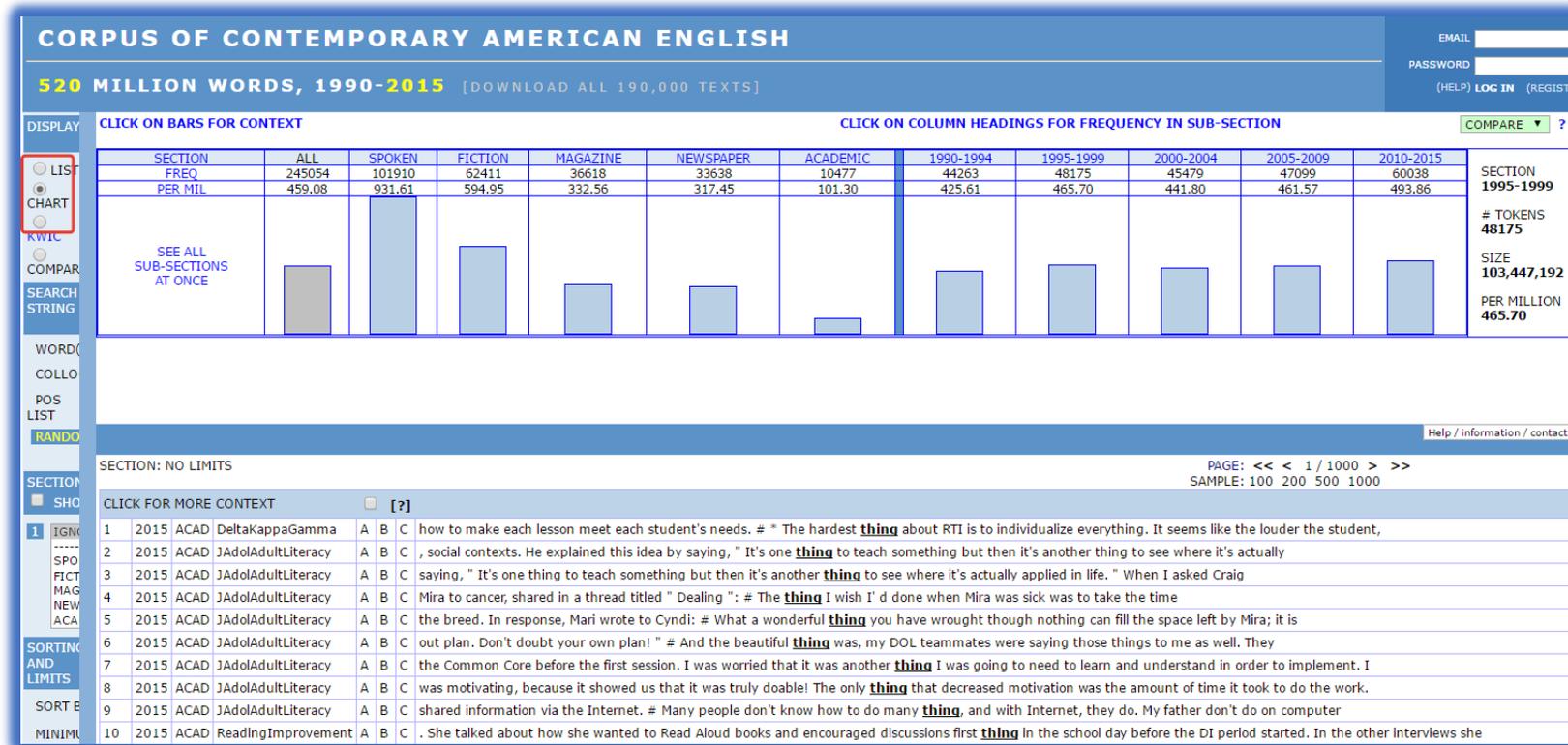
Date	2015
Publication information	Spring2014, Vol. 80 Issue 3, p11-23. 13p.
Title	Understanding Secondary Teachers' Concerns about RTI: Purposeful Professional Communication
Author	Isbell, Laura J.; Szabo, Susan;
Source	Delta Kappa Gamma Bulletin

Expanded context:

was a concern to them, as they did not understand what to document or how to document student RTI strategies used in the classroom. # * Should I document everything? # * Do I need to turn some form of documentation in to the RTI specialist or to the administrators? # * Where do I keep everything? Do I keep it or do I need to turn it in to someone? Six teachers indicated individualization as a concern. They were uncertain about how to make each lesson meet each student's needs. # * The hardest thing about RTI is to individualize everything. It seems like the louder the student, the more help they receive. I have such a large class and it is hard to work with everyone. # * Making sure the right kids receive the right services and strategies is a concern because sometimes kids who don't need services are on it. # * I would like to help all the kids and find what benefits each student to be successful. # Goal statements. When participants were asked where they thought

[Return to KWIC entries](#)

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**.

CORPUS OF CONTEMPORARY AMERICAN ENGLISH

520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

EMAIL PASSWORD (HELP) LOG IN (RE)

DISPLAY **CLICK ON BARS FOR CONTEXT** **CLICK ON COLUMN HEADINGS FOR FREQUENCY IN SUB-SECTION** COMPARE ▾

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2015
FREQ	245054	101910	62411	36618	33638	10477	44263	48175	45479	47099	60038
PER MIL	459.08	931.61	594.95	332.56	317.45	101.30	425.61	465.70	441.80	461.57	493.86

SEE ALL SUB-SECTIONS AT ONCE

SECTION SPOKEN
TOKENS 101910
SIZE 109,391,6
PER MILLI 931.61

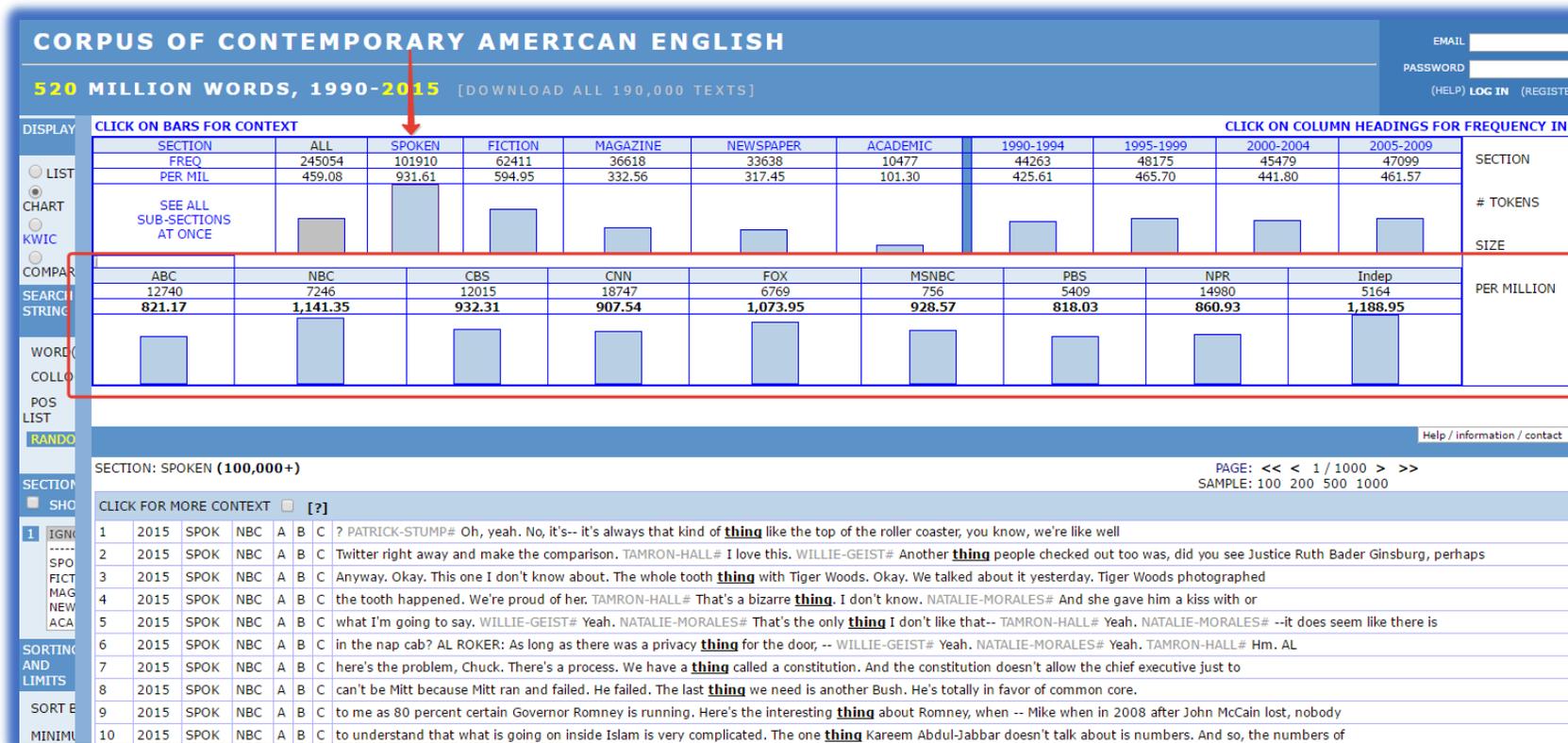
KEYWORD IN CONTEXT DISPLAY Help / information / con

SECTION: SPOKEN (100,000+) PAGE: << < 1 / 1000 > >> SAMPLE: 100 200 500 1000

CLICK FOR MORE CONTEXT [?]

1	IGN	2015	SPOK	NBC	A	B	C	? PATRICK-STUMP#	Oh, yeah. No, it's-- it's always that kind of thing like the top of the roller coaster, you know, we're like well										
2	SPO	2015	SPOK	NBC	A	B	C	Twitter right away and make the comparison. TAMRON-HALL#	I love this. WILLIE-GEIST# Another thing people checked out too was, did you see Justice Ruth Bader Ginsburg, perhaps										
3	FICT	2015	SPOK	NBC	A	B	C	Anyway. Okay. This one I don't know about. The whole tooth thing with Tiger Woods. Okay. We talked about it yesterday. Tiger Woods photographed											
4	MAG	2015	SPOK	NBC	A	B	C	the tooth happened. We're proud of her. TAMRON-HALL#	That's a bizarre thing . I don't know. NATALIE-MORALES# And she gave him a kiss with or										
5	NEW	2015	SPOK	NBC	A	B	C	what I'm going to say. WILLIE-GEIST#	Yeah. NATALIE-MORALES# That's the only thing I don't like that-- TAMRON-HALL#	Yeah. NATALIE-MORALES# --it does seem like there is									
6	ACA	2015	SPOK	NBC	A	B	C	in the nap cab? AL ROKER: As long as there was a privacy thing for the door, -- WILLIE-GEIST#	Yeah. NATALIE-MORALES#	Yeah. TAMRON-HALL#	Hm. AL								
7		2015	SPOK	NBC	A	B	C	here's the problem, Chuck. There's a process. We have a thing called a constitution. And the constitution doesn't allow the chief executive just to											
8		2015	SPOK	NBC	A	B	C	can't be Mitt because Mitt ran and failed. He failed. The last thing we need is another Bush. He's totally in favor of common core.											
9		2015	SPOK	NBC	A	B	C	to me as 80 percent certain Governor Romney is running. Here's the interesting thing about Romney, when -- Mike when in 2008 after John McCain lost, nobody											
10		2015	SPOK	NBC	A	B	C	to understand that what is going on inside Islam is very complicated. The one thing Kareem Abdul-Jabbar doesn't talk about is numbers. And so, the numbers of											

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*).

CORPUS OF CONTEMPORARY AMERICAN ENGLISH

520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

EMAIL

PASSWORD

(HELP) LOG IN (RE)

DISPLAY

CONCORDANCING, with re-sorting and part of speech highlighting ([more information](#))

KEYWORD IN CONTEXT DISPLAY [Help / information / co](#)

100,000+ TOKENS L - - - 1 2 3 R * |

CLICK FOR MORE CONTEXT [?]

1	2008	NEWS	USAToday	A	B	C	who does not wear a knee brace ; says the toughest thing about the surgery is answering the same question over and over .
2	1992	SPOK	CBS_Street	A	B	C	she woke up . And there was fire that was -- thing all over the household and our ... (unintelligible) so she
3	2008	FIC	Analog	A	B	C	is , if there is a god-and there is such a thing as a soul . " Paul smiled wistfully . " Who knows
4	2012	SPOK	NPR	A	B	C	these databases . NEAL-CONAN# Ozzie Nelson , is there such a thing as too much information , too much data ? RICK-NELSON# Again ,
5	1997	FIC	Bk:PlumIsland	A	B	C	I was going to see it today , but this other thing came up . " We 're only open weekends after Labor Day
6	1996	SPOK	CNN_News	A	B	C	we found out about it to ensure that this kind of thing can never happen again . The president apologized to the
7	2012	FIC	Analog	A	B	C	# " Why ? " Karen asked . # " This thing can stop time , or entropy . I ca n't imagine how
8	1995	MAG	USNWR	A	B	C	would hope if I were elected and we had such a thing come up again , I could point with some pride to the
9	2001	SPOK	CBS_FaceNation	A	B	C	of this bill should be declared unconstitutional , the whole thing dies . ! MCCAIN : Which is almost never done
10	2013	FIC	Bk:SubtleBodies	A	B	C	thought he was n't sharing . Then he might not some thing down on a scrap of paper or he might not . The
11	1990	FIC	Ploughshares	A	B	C	it didn't. on the way out he 'd say the same thing every time he saw Roy Magoon asleep on a battered recliner in
12	2002	FIC	NewEnglandRev	A	B	C	push my hand away . # " Charly 's the best thing for a nightmare , " I 'd remind her . " Abuelita
13	2015	SPOK	ABC	A	B	C	one issue to the Democrats . And that 's a good thing for the country . JAMES-CARVILLE# Look , what 's going to
14	2009	MAG	MensHealth	A	B	C	, your computer will sit there doing its inscrutable thing for up to 6 hours , says Ken Colburn , president of
15	2000	ACAD	AmerIndianQ	A	B	C	that person . So when that person is gone , this thing has n't got anything to eat or a person go to ,
16	2011	NEWS	USAToday	A	B	C	get done ? Compromise . # " The most important key thing here is our fans and the support from the people and the
17	2009	SPOK	ABC_PrimeTime	A	B	C	there are interested in seeing this happen . The most important thing I can say , James , on this issue is , if
18	1996	SPOK	NPR_Weekend	A	B	C	decisions and say , ' Yeah , I did the best thing I could yesterday , now how do I approach it again today
19	1990	FIC	Ploughshares	A	B	C	and I thought , Hands , and I remembered the only thing I ever remembered about my mother , how she cooled me down
20	2006	SPOK	Ind_Oprah	A	B	C	her job started taking a toll . Ms-GOLDRICK : One sad thing I found out about nursing was that a lot of elderly people
21	2008	FIC	Bk:GorillaBlack	A	B	C	Gorilla Black ! " I yelled out ! // The next thing I knew everybody took off running as if they had seen
22	1996	NEWS	Atlanta	A	B	C	He 's a better passer than I thought . And the thing I like about him so much : He knows what he has

SEARCH STRING

WORD(

COLLO

POS LIST

RANDO

SECTION

SHO

IGNI

SPO

FICT

MAG

NEW

ACA

DISPLAY / SORT

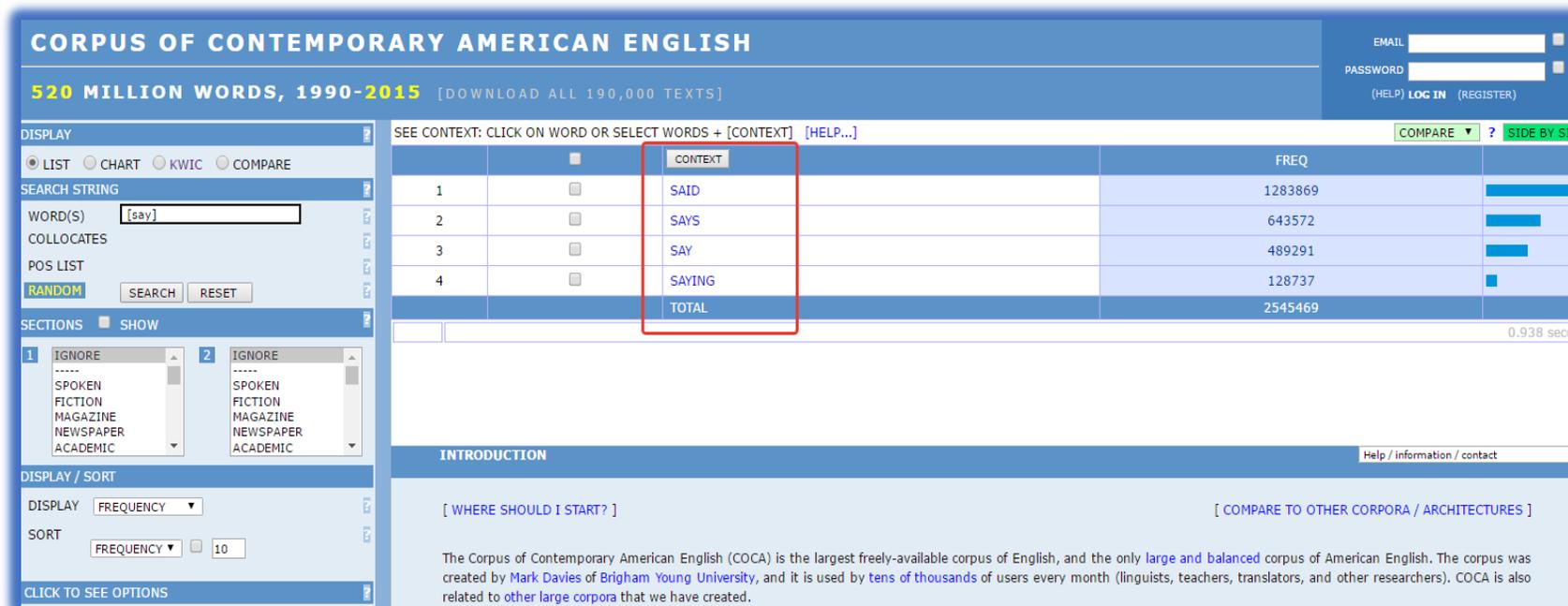
DISPLA

SORT

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*.



CORPUS OF CONTEMPORARY AMERICAN ENGLISH

520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

EMAIL
PASSWORD
(HELP) [LOG IN](#) (REGISTER)

DISPLAY: LIST CHART KWIC COMPARE

SEARCH STRING:

WORD(S):

COLLOCATES:

POS LIST:

[RANDOM](#)

SECTIONS SHOW

1 IGNORE
.....
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC

2 IGNORE
.....
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC

DISPLAY / SORT

DISPLAY:

SORT: 10

[CLICK TO SEE OPTIONS](#)

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

	CONTEXT	FREQ	
1	SAID	1283869	
2	SAYS	643572	
3	SAY	489291	
4	SAYING	128737	
	TOTAL	2545469	

0.938 seconds

INTRODUCTION [Help / information / contact](#)

[[WHERE SHOULD I START?](#)] [[COMPARE TO OTHER CORPORA / ARCHITECTURES](#)]

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, and the only large and balanced corpus of American English. The corpus was created by Mark Davies of Brigham Young University, and it is used by tens of thousands of users every month (linguists, teachers, translators, and other researchers). COCA is also related to other large corpora that we have created.

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.

CORPUS OF CONTEMPORARY AMERICAN ENGLISH
520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

SEARCH STRING: [=beautiful]

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

		CONTEXT	FREQ	
1	<input type="checkbox"/>	BEAUTIFUL [S]	54025	
2	<input type="checkbox"/>	WONDERFUL [S]	33687	
3	<input type="checkbox"/>	ATTRACTIVE [S]	14024	
4	<input type="checkbox"/>	STRIKING [S]	12407	
5	<input type="checkbox"/>	LOVELY [S]	12193	
6	<input type="checkbox"/>	HANDSOME [S]	9806	
7	<input type="checkbox"/>	STUNNING [S]	6966	
8	<input type="checkbox"/>	GORGEOUS [S]	6503	
9	<input type="checkbox"/>	CHARMING [S]	6267	

INTRODUCTION [WHERE SHOULD I START?] [COMPARE TO OTHER CORPORA / ARCHITECTURES]

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, and the only large and balanced corpus of American English. The corpus was created by Mark Davies of Brigham Young University, and it is used by tens of thousands of users every month (linguists, teachers, translators, and other researchers). COCA is also related to other large corpora that we have created.

Поиск по любому из заданных слов

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

Например: *cold/cool/frosty*

CORPUS OF CONTEMPORARY AMERICAN ENGLISH
520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

EMAIL
PASSWORD
(HELP) [LOG IN](#) (REGISTER)

DISPLAY: LIST CHART KWIC COMPARE

SEARCH STRING: WORD(S)

COLLOCATES: POS LIST:

SECTIONS: 1 IGNORE 2 IGNORE

SORTING AND LIMITS: SORTING: FREQUENCY MINIMUM: FREQUENCY

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

		CONTEXT	FREQ	
1	<input type="checkbox"/>	COLD	58238	<div style="width: 58%;"></div>
2	<input type="checkbox"/>	COOL	36532	<div style="width: 36%;"></div>
3	<input type="checkbox"/>	FROSTY	870	<div style="width: 0.8%;"></div>
		TOTAL	95640	

0.324 seconds

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.



CORPUS OF CONTEMPORARY AMERICAN ENGLISH
520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

ACCESS: 1, [history | lists | profile | log]

DISPLAY: LIST CHART KWIC COMPARE

SEARCH STRING: WORD(S) COLLOCATES POS LIST [RANDOM] [SEARCH] [RESET]

SECTIONS: SHOW

1 IGNORE 2 IGNORE

SORTING AND LIMITS: SORTING: FREQUENCY MINIMUM: FREQUENCY 10

CLICK TO SEE OPTIONS

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

		CONTEXT	FREQ	
1	<input type="checkbox"/>	SERIOUS	57193	
2	<input type="checkbox"/>	VARIOUS	52294	
3	<input type="checkbox"/>	RELIGIOUS	51913	
4	<input type="checkbox"/>	PREVIOUS	36230	
5	<input type="checkbox"/>	DANGEROUS	28852	
6	<input type="checkbox"/>	FAMOUS	28179	
7	<input type="checkbox"/>	OBVIOUS	23721	
8	<input type="checkbox"/>	ENORMOUS	19524	
9	<input type="checkbox"/>	NUMEROUS	17504	
10	<input type="checkbox"/>	NERVOUS	16026	
11	<input type="checkbox"/>	TREMENDOUS	12821	
12	<input type="checkbox"/>	CURIOUS	11269	
13	<input type="checkbox"/>	INDIGENOUS	10138	
14	<input type="checkbox"/>	CONTINUOUS	9937	

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 (.)*

CORPUS OF CONTEMPORARY AMERICAN ENGLISH
520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

SEARCH STRING: chair.{nn*}
POS LIST: noun.ALL

	CONTEXT	FREQ
1	CHAIR	42635

KEYWORD IN CONTEXT DISPLAY

SECTION: NO LIMITS

PAGE: << < 1 / 427 > >>
SAMPLE: 100 200 500 1000

CLICK FOR MORE CONTEXT	[?]	
1	2015 NEWS WashPost	A B C credit Woolley's team with making the difference. McCormick, who served as DNC chair from 1916 to 1919, said Woolley had put together " the st
2	2015 NEWS WashPost	A B C had been kept on, though others handled the duties - Mrs. Brady was named chair of Handgun Control. In 1991, she became the chair of Handg
3	2015 NEWS WashPost	A B C Mrs. Brady was named chair of Handgun Control. In 1991, she became the chair of Handgun Control's sister organization, the Center to Prevent H
4	2015 NEWS WashPost	A B C . " Agriculture is already taking a hard hit, " said Felicia Marcus, chair of the State Water Resources Control Board. She called the 80 percent to 20
5	2015 NEWS WashPost	A B C living room; the modular, Knoll-like sofas; the Lied Mobler black leather lounge chair ; the built-in walnut cabinetry; the countertop cocktail bar; th
6	2015 NEWS WashPost	A B C as well be my living room. The wall of windows. The Eames Time-Life chair . The Florence Knoll settee. The Paul McCobb coffee table. The Lightolier
7	2015 NEWS WashPost	A B C 'd begun to collect mid-century classics: a Knoll dresser here, an Ib Kofod-Larsen chair there. That was late 2008 and early 2009, the height of " f
8	2015 NEWS WashPost	A B C percent in North America over the past seven years; sales of the Eames Time-Life chair , which is prominently featured in the SCDP conference roo
9	2015 NEWS WashPost	A B C Claudette Didul goes to extreme lengths to ensure that everything - the Poul Volther Corona chair in Roger Sterling's all-white office; the boxy Kn
10	2015 NEWS WashPost	A B C organic quality of nature itself? " The solution could be an Alvar Aalto Hallway chair from 1932. Or it could be a Jasper Morrison Air-Chair from 199

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

CORPUS OF CONTEMPORARY AMERICAN ENGLISH

520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

EMAIL
PASSWORD
(HELP) [LOG IN](#) (REGISTER)

DISPLAY: LIST CHART KWIC COMPARE

SEARCH STRING:

WORD(S):

COLLOCATES

POS LIST:

[RANDOM](#) [SEARCH](#) [RESET](#)

SECTIONS SHOW

1 IGNORE
.....
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC

2 IGNORE
.....
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC

DISPLAY / SORT

DISPLAY:

SORT: 10

CLICK TO SEE OPTIONS

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

COMPARE ? SIDE BY SIDE

	<input type="checkbox"/>	CONTEXT	FREQ	
1	<input type="checkbox"/>	CHAIR BACK	235	
2	<input type="checkbox"/>	CHAIR LEGS	78	
3	<input type="checkbox"/>	CHAIR LIFT	52	
4	<input type="checkbox"/>	CHAIR LEG	47	
5	<input type="checkbox"/>	CHAIR ARM	45	
6	<input type="checkbox"/>	CHAIR RISE	44	
7	<input type="checkbox"/>	CHAIR SEAT	43	
8	<input type="checkbox"/>	CHAIR RAIL	32	
9	<input type="checkbox"/>	CHAIR ARMS	29	

KEYWORD IN CONTEXT DISPLAY [Help / information / contact](#)

SECTION: NO LIMITS

PAGE: << < 1 / 427 > >>
SAMPLE: 100 200 500 1000

CLICK FOR MORE CONTEXT [?]

1	2015	NEWS	WashPost	A B C	credit Woolley's team with making the difference. McCormick, who served as DNC chair from 1916 to 1919, said Woolley had put together " the st
2	2015	NEWS	WashPost	A B C	had been kept on, though others handled the duties - Mrs. Brady was named chair of Handgun Control. In 1991, she became the chair of Handgu
3	2015	NEWS	WashPost	A B C	Mrs. Brady was named chair of Handgun Control. In 1991, she became the chair of Handgun Control's sister organization, the Center to Prevent H
4	2015	NEWS	WashPost	A B C	." Agriculture is already taking a hard hit," said Felicia Marcus, chair of the State Water Resources Control Board. She called the 80 percent to 20
5	2015	NEWS	WashPost	A B C	living room; the modular, Knoll-like sofas; the Lied Mobler black leather lounge chair ; the built-in walnut cabinetry; the countertop cocktail bar; the
6	2015	NEWS	WashPost	A B C	as well be my living room. The wall of windows. The Eames Time-Life chair . The Florence Knoll settee. The Paul McCobb coffee table. The Lightolier
7	2015	NEWS	WashPost	A B C	'd begun to collect mid-century classics: a Knoll dresser here, an Ib Kofod-Larsen chair there. That was late 2008 and early 2009, the height of " M
8	2015	NEWS	WashPost	A B C	percent in North America over the past seven years; sales of the Eames Time-Life chair , which is prominently featured in the SCDP conference room
9	2015	NEWS	WashPost	A B C	Claudette Didul goes to extreme lengths to ensure that everything - the Poul Volther Corona chair in Roger Sterling's all-white office; the boxy Kno
10	2015	NEWS	WashPost	A B C	organic quality of nature itself? " The solution could be an Alvar Aalto Hallway chair from 1932. Or it could be a Jasper Morrison Air-Chair from 199

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.

The screenshot shows the search interface for the Corpus of Contemporary American English. The search string is "chair [nn*]". The POS list is set to "noun.ALL". The collocates are displayed in a table with columns for rank, context, frequency, and various statistical measures.

	CONTEXT	FREQ	ALL	%	MI
1	BACK	235	655892	0.04	6.95
2	LEGS	78	37583	0.21	9.48
3	LIFT	48	17109	0.28	9.92
4	LEG	47	24587	0.19	9.37
5	ARM	45	48214	0.09	8.33
6	RISE	44	38457	0.11	8.62
7	SEAT	43	40723	0.11	8.51
8	RAIL	32	9331	0.34	10.21
9	ARMS	29	62106	0.05	7.33
10	RIGHT	23	589157	0.00	3.75
11	LIFTS	19	5482	0.35	10.22
12	BACKS	18	9457	0.19	9.36
13	SEATS	18	18065	0.10	8.43
14	IMPIRE	17	1017	1.67	12.49

Below the table, there is a section for "KEYWORD IN CONTEXT DISPLAY" showing example sentences with the word "chair" highlighted in context.

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*).

CORPUS OF CONTEMPORARY AMERICAN ENGLISH
520 MILLION WORDS, 1990-2015 [DOWNLOAD ALL 190,000 TEXTS]

SEARCH STRINGS: WORD(S) morning evening; COLLOCATES [*]; POS LIST adj.ALL

SEE CONTEXT: CLICK ON NUMBERS (WORD 1 OR 2)

WORD 1 (W1): MORNING (3.00)

WORD	W1	W2	W1/W2	SCORE
1 IN-DEPTH	19	0	38.0	12.7
2 WEE	19	0	38.0	12.7
3 SUNNY	269	8	33.6	11.2
4 TOP	134	4	33.5	11.2
5 MID	32	1	32.0	10.7
6 ROUNDTABLE	16	0	32.0	10.7
7 EARLIEST	15	0	30.0	10.0
8 BRAVER	14	0	28.0	9.3
9 HIGHEST	14	0	28.0	9.3
10 MASSIVE	14	0	28.0	9.3
11 PROUD	14	0	28.0	9.3

WORD 2 (W2): EVENING (0.33)

WORD	W2	W1	W2/W1	SCORE
1 ROMANTIC	57	0	114.0	341.9
2 ENCHANTED	66	1	66.0	198.0
3 DARKENING	16	0	32.0	96.0
4 ENTERTAINING	14	0	28.0	84.0
5 STRAPLESS	13	0	26.0	78.0
6 BEADED	12	0	24.0	72.0
7 FADING	11	0	22.0	66.0
8 ENJOYABLE	10	0	20.0	60.0
9 FANCY	9	0	18.0	54.0
10 STILLWATER	9	0	18.0	54.0
11 LOW-CUT	8	0	16.0	48.0

KEYWORD IN CONTEXT DISPLAY