Экспертиза usability

News of the Web Corpus

Проектная работа по Компьютерным инструментам лингвистического исследования, 1 курс, Фундаментальная и компьютерная лингвистика

Подготовили:
Ветошкина Арина (E-mai)(avvetoshkina@edu.se.ru);
Гладышевская Александра (E-mail)(amgladyshevskaya@edu.hse.ru);
Павлова Марина (E-mail)(mkpavlova@edu.hse.ru);
Пономарева Полина (E-mail)(paponomareva_2@edy.hse.ru);

1 модуль, 2017-2018 учебный год.

News of the Web Corpus **-** корпус, созданный американским лингвистом Марком Дэвисом. Содержит около 5 миллиардов слов и 6 миллионов текстов. Такой объем объясняется системой сбора данных для корпуса: он пополняется статьями интернет-изданий 20-и англоговорящих стран: США, Канады, Великобритании, Ирландии, Австралии, Новой Зеландии, Индии, Шри-Ланки, Пакистана, Бангладеша, Малайзии, Сингапура, Филиппин, Гонконга, ЮАР, Нигерии, Ганы, Кении, Танзании и Ямайки. Ежедневно корпус пополняется на 4-5 миллионов слов. Пополнение и аннотирование корпуса происходит автоматически (с помощью следующих программ: jusTex, CLAWS PoS tagger for English**),** а из-за большого объема данных проверить вручную точность разметки сложно, в нем встречаются нелепые слова, попавшие в него по ошибке. Тем не менее корпус чрезвычайно полезен тем, кто изучает современное состояние английского языка (не только в Британии или США, но и в других англоговорящих странах, что очень важно) и фиксирует происходящие в нем изменения, а также социологам, политологам, журналистам, школьникам и студентам (для улучшения уровня владения английским языком).

Дизайн:

Первое, на что пользователь корпуса обращает внимание - это дизайн. Корпус приобрел новый интерфейс еще в 2016 году (рис. 1).

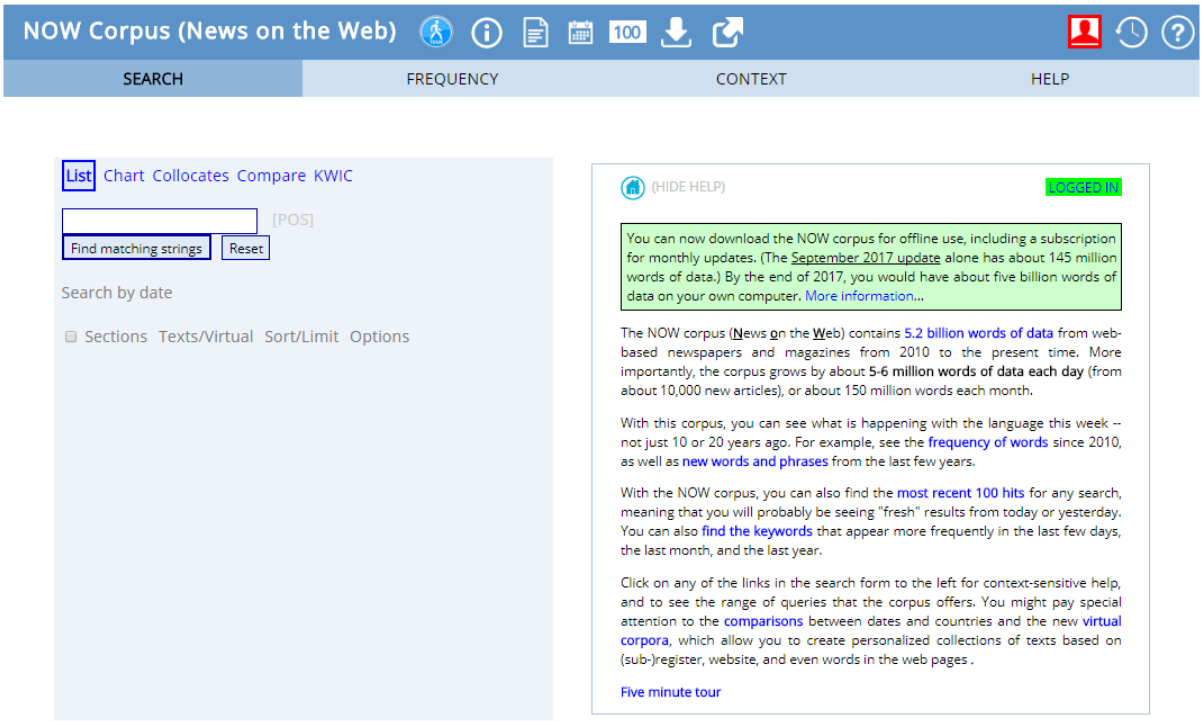


рис. 1

Дизайн корпуса выполнен в спокойных тонах. Чаще всего использованы голубой, синий и их оттенки. Отдельные элементы, блоки выделены желтым, красным, зеленым. В целом корпус довольно гармоничен, нет отвлекающих

элементов, а важные выделены (рис.1). Новый интерфейс корпуса также адаптирован к любым видам электронных устройств и работает стабильно как на ПК и планшетах, так и мобильных устройствах (смартфонах), однако использование корпуса на последних затруднено ввиду мелкого шрифта. (рис.2).

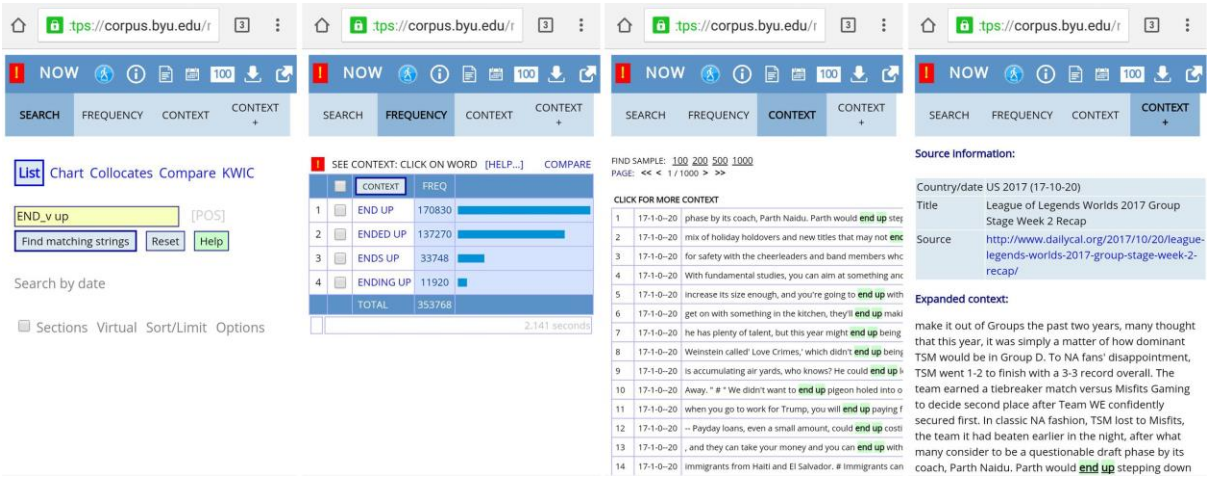


рис.2

Вверху страницы расположены основные ссылки: 5-minute tour, информация, статистика пополнения по месяцам и т.д. Значки могут показаться непонятными, но при наведении всплывает их значение (рис.3), а во вкладке HELP можно найти подробное описание функций каждой из них.

рис.3

Кнопки в разделе поиска удобно и последовательно расположены (рис.4). Есть так называемые «серые» кнопки, при нажатии на которые всплывают элементы для более подробного поиска с доступной и понятной инструкцией справа, что значительно облегчает работу в данном корпусе (об этом в разделе “Onboarding”).

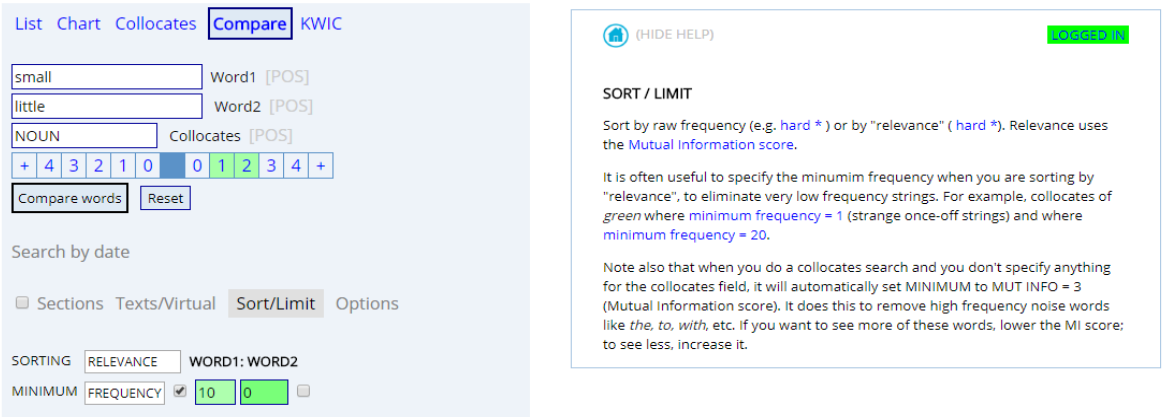


рис.4 *(форма поиска, нажатая «серая» кнопка Sort/Limit, активные подсказки)*

Результаты поиска на странице CONTEXT сведены в удобную таблицу, в которо-й искомое слово/фраза/выражение выделено определенным цветом (рис.5). Шрифт небольшой, это немного затрудняет чтение контекстов и ориентирование по таблице.

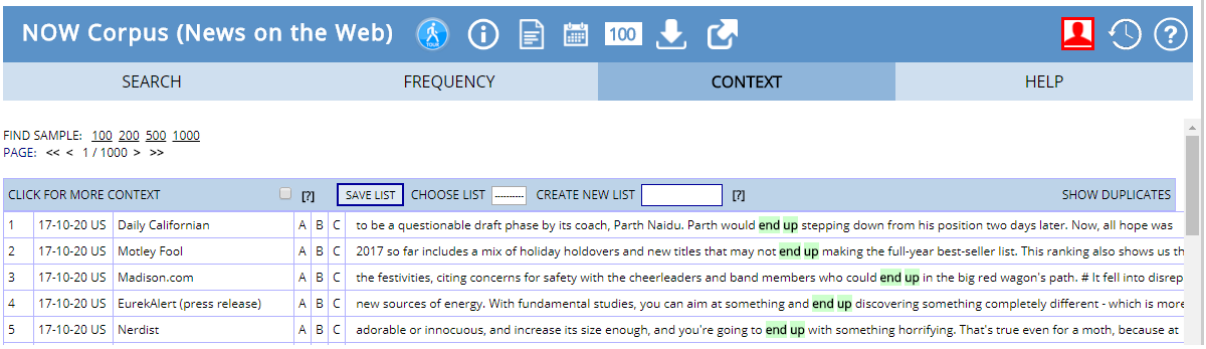


рис.5

Большим недостатком является отсутствие какой-либо графики, которая бы показывала процесс/статус обработки поискового запроса и выдачи данных. Да, время, потраченное на поиск, указывается в самом низу выдачи, однако иногда непонятно, откликнулся ли сайт на запрос пользователя: для этого нет никакого сигнала (например, всплывающего окна или другого показателя выполнения – крутящегося кружочка и т.п.).

В целом, интерфейс приятный на вид и довольно удобный, однако мелкие недочеты, которые имеются (мелкий шрифт, отсутствие показателя отклика) доставляют неудобства, но не затрудняют работу с корпусом.

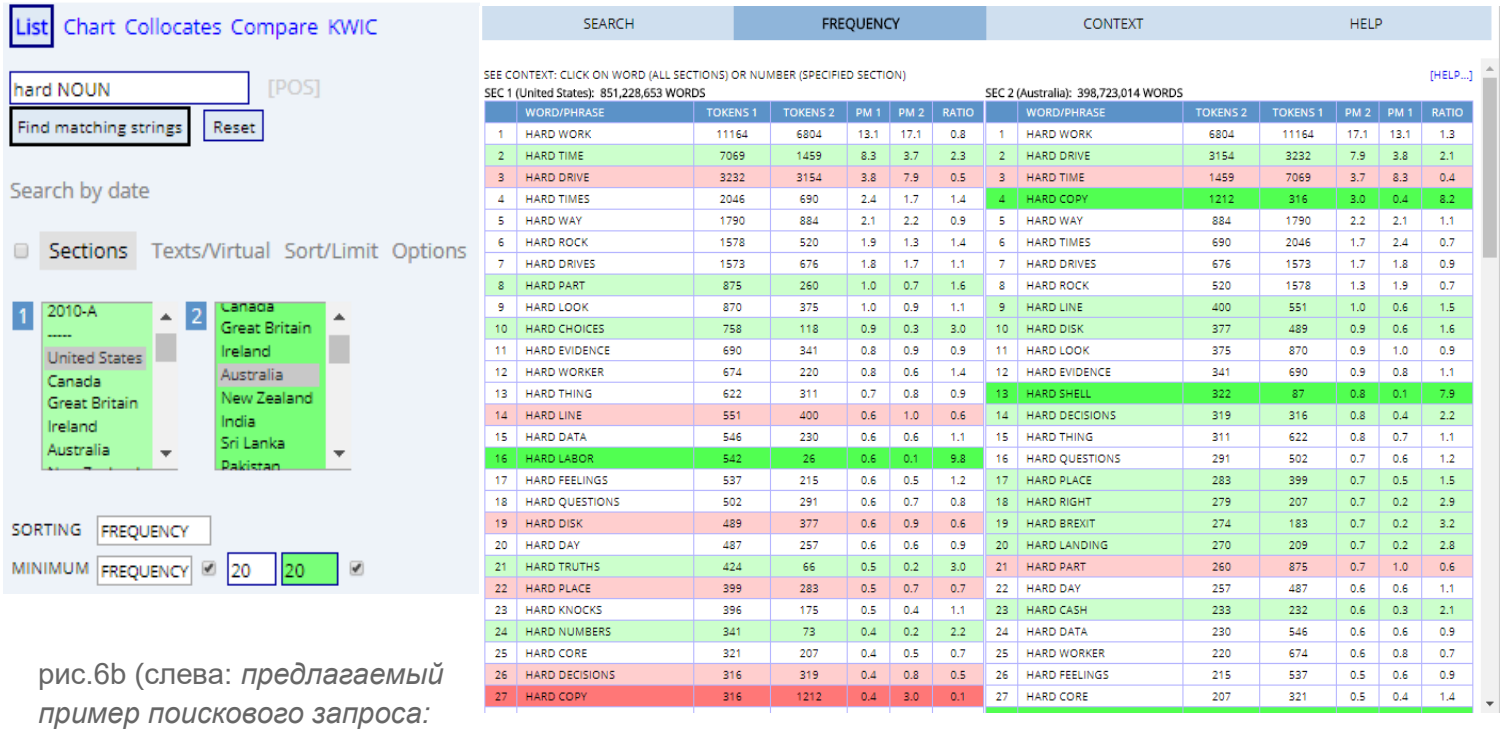
Onboarding (ресурс глазами новичка) и инструкции:

Ресурс появляется на первой строке выдачи результатов запроса в поисковиках. Адрес корпуса в Интернете достаточно короткий и легкий для запоминания, что, на наш взгляд, является плюсом. Форма поиска находится на главной странице корпуса (рис.1) - туда можно попасть как нажав на название ресурса в левом верхнем углу, так и войдя во вкладку “Search”, что есть логично и удобно.

Со страниц с результатами поиска можно легко перейти на интернет издания, откуда была взята информация. Эта функция позволяет пользователям сравнить информацию на порталах и в разных странах, особенно она актуальна для политологов, социологов, так как позволяет увидеть картину мира.

Подсказка, автоматически появляющаяся в правой части экрана, коротко описывает часть функций корпуса: от возможных критериев поиска до сторонних функций (например, предлагается список неологизмов, составленный WordSpy.com). Удобно, что перечисленные функции являются активными ссылками на примеры возможных поисковых запросов: это помогает разобраться и с функционалом, и с синтаксисом запросов. Другие функции и примеры запросов (рис.6b, 6c) приведены в разделе “Five minute tour”, в который

также легко попасть: ссылка есть и в самой подсказке, и в “шапке” сайта (крайний левый значок; рис. 3).



	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	HARD WORK	11164	6804	13.1	17.1	0.8
2	HARD TIME	7069	1459	8.3	3.7	2.3
3	HARD DRIVE	3232	3154	3.8	7.9	0.5
4	HARD TIMES	2046	690	2.4	1.7	1.4
5	HARD WAY	1790	884	2.1	2.2	0.9
6	HARD ROCK	1578	520	1.9	1.3	1.4
7	HARD DRIVES	1573	676	1.8	1.7	1.1
8	HARD PART	875	260	1.0	0.7	1.6
9	HARD LOOK	870	375	1.0	0.9	1.1
10	HARD CHOICES	758	118	0.9	0.3	3.0
11	HARD EVIDENCE	690	341	0.8	0.9	0.9
12	HARD WORKER	674	220	0.8	0.6	1.4
13	HARD THING	622	311	0.7	0.8	0.9
14	HARD LINE	551	400	0.6	1.0	0.6
15	HARD DATA	546	230	0.6	0.6	1.1
16	HARD LABOR	542	26	0.6	0.1	9.8
17	HARD FEELINGS	537	215	0.6	0.5	1.2
18	HARD QUESTIONS	502	291	0.6	0.7	0.8
19	HARD DISK	489	377	0.6	0.9	0.6
20	HARD DAY	487	257	0.6	0.6	0.9
21	HARD TRUTHS	424	66	0.5	0.2	3.0
22	HARD PLACE	399	283	0.5	0.7	0.7
23	HARD KNOCKS	396	175	0.5	0.4	1.1
24	HARD NUMBERS	341	73	0.4	0.2	2.2
25	HARD CORE	321	207	0.4	0.5	0.7
26	HARD DECISIONS	316	319	0.4	0.8	0.5
27	HARD COPY	316	1212	0.4	3.0	0.1

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	HARD WORK	6804	11164	17.1	13.1	1.3
2	HARD DRIVE	3154	3232	7.9	3.8	2.1
3	HARD TIME	1459	7069	3.7	8.3	0.4
4	HARD COPY	1212	316	3.0	0.4	8.2
5	HARD WAY	884	1790	2.2	2.1	1.1
6	HARD TIMES	690	2046	1.7	2.4	0.7
7	HARD DRIVES	676	1573	1.7	1.8	0.9
8	HARD ROCK	520	1578	1.3	1.9	0.7
9	HARD LINE	400	551	1.0	0.6	1.5
10	HARD DISK	377	489	0.9	0.6	1.6
11	HARD LOOK	375	870	0.9	1.0	0.9
12	HARD EVIDENCE	341	690	0.9	0.8	1.1
13	HARD SHELL	322	87	0.8	0.1	7.9
14	HARD DECISIONS	319	316	0.8	0.4	2.2
15	HARD THING	311	622	0.8	0.7	1.1
16	HARD QUESTIONS	291	502	0.7	0.6	1.2
17	HARD PLACE	283	399	0.7	0.5	1.5
18	HARD RIGHT	279	207	0.7	0.2	2.9
19	HARD BREXIT	274	183	0.7	0.2	3.2
20	HARD LANDING	270	209	0.7	0.2	2.8
21	HARD PART	260	875	0.7	1.0	0.6
22	HARD DAY	257	487	0.6	0.6	1.1
23	HARD CASH	233	232	0.6	0.3	2.1
24	HARD DATA	230	546	0.6	0.6	0.9
25	HARD WORKER	220	674	0.6	0.8	0.7
26	HARD FEELINGS	215	537	0.5	0.6	0.9
27	HARD CORE	207	321	0.5	0.4	1.4

рис.6b (слева: *предлагаемый пример поискового запроса: сравнить словосочетания “HARD + NOUN” в американский и австралийских онлайн-издания)*
рис.6c *(справа: выдача результатов по примеру HARD + NOUN US vs. AU)*

Итак, как можно увидеть из приложенных нами скриншотов, примеры даны нетривиальные, это дает возможность сразу познакомиться с широким функционалом корпуса. Нам, как новичкам, разобраться и понять структуру корпуса и поиска по нему было нетрудно. Ссылки на помощь в использовании корпуса доступны как в “шапке” сайта (рис.7), так и на каждой отдельной вкладке непосредственно для работы с выдачей результатов того или иного запроса.

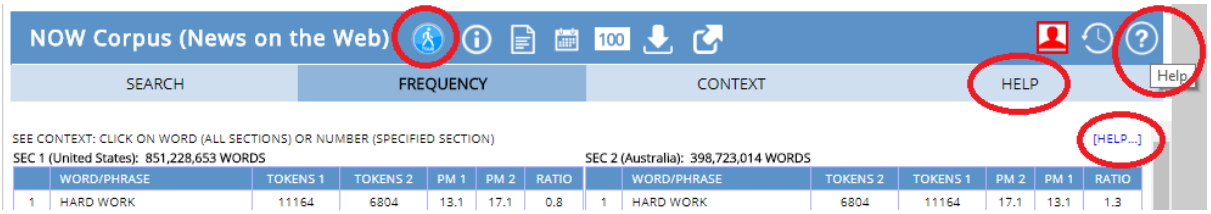


рис.7

Несомненным достоинством является интерактивная система подсказок (рис.4): они появляются в зависимости от действий пользователя. Например, при наведении курсора на различные разделы поисковой формы появляются подсказки, касающиеся использования именно этой конкретной функции поиска.

Там же, в подсказках, имеется глоссарий - нужная вещь для корпуса, с которым могут не только лингвисты (но и, например, социологи, политологи, школьники).

На сторонних сайтах нет примеров использования NOW Corpus (ни видео-уроков, ни статей), однако есть тьюториалы для других корпусов Марка Дэвиса,

работающих на аналогичной NOW Corpus'у платформе (различны лишь базы текстов и функции). Например, имеется достаточное количество видео-уроков для Davies, Mark. (2015) *The Wikipedia Corpus.*

Однако надо отметить, что подсказки на страницах News of the Web Corpus даны в достаточном количестве, что позволяет новичку и без тьюториалов быстро и без проблем разобраться в навигации по корпусу и функциях поиска.

Слои разметки:

1. Мета-разметка:
   - Сведения о стране и дата добавления (Country/date)
   - Заголовок (Title)
   - Ресурс (Source)
2. Морфологическая разметка:
   - Part of speech coloring:

noun pronoun proper noun adjective verb adverb preposition

Отсутствуют:

- семантическая
- синтаксическая

Продвинутый функционал:

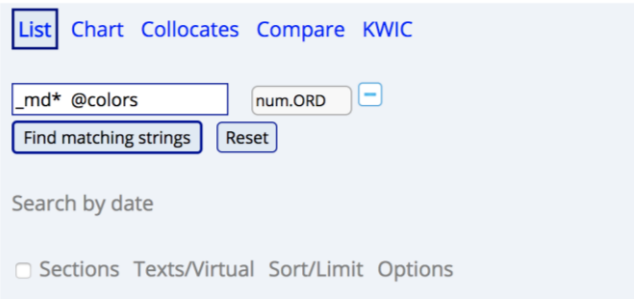
Корпус позволяет производить поиск в различных режимах. Может использоваться для поиска:

a. отдельных слов
b. различных форм слова
c. слов, подходящих под заданные формулы
d. фраз и словосочетаний
e. синонимов
f. слов из созданных пользователем списков
g. частей речи

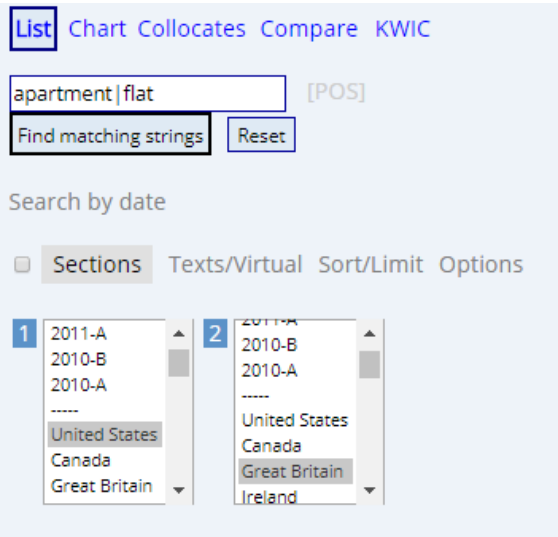
Для любого режима поиска можно указать дополнительные настройки: настройки сортировки, выбор сектора (дата или страна), выбор подкорпуса (который можно создать вручную залогинившемуся пользователю - это может быть любая необходимая ему выборка, сохраненная в личном кабинете).

- List: Показывает частотность конкретного запроса во всем корпусе.

Пример запроса: Найдем словосочетания, состоящие из порядкового числительного и любого слова, обозначающего название цвета (список @colors был создан вручную до выполнения запроса).



Как мы можем заметить, выдача отсортирована по частоте вхождения данного запроса. Для каждого результата мы можем открыть контекст употребления с указанием источника (и активной ссылкой на статью) и другой информации во вкладке “CONTEXT”.



Пример: сравним, как количественно различается употребление слов “apartment” и “flat” в США и Великобритании (American English vs British English.
Контексты, аналогично, можно увидеть во вкладке CONTEXT

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

SEC 1 (United States): 851,228,653 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	APARTMENT	49895	19075	58.6	27.3	2.1
2	FLAT	24967	44180	29.3	63.2	0.5

SEC 2 (Great Britain): 698,598,367 WORDS

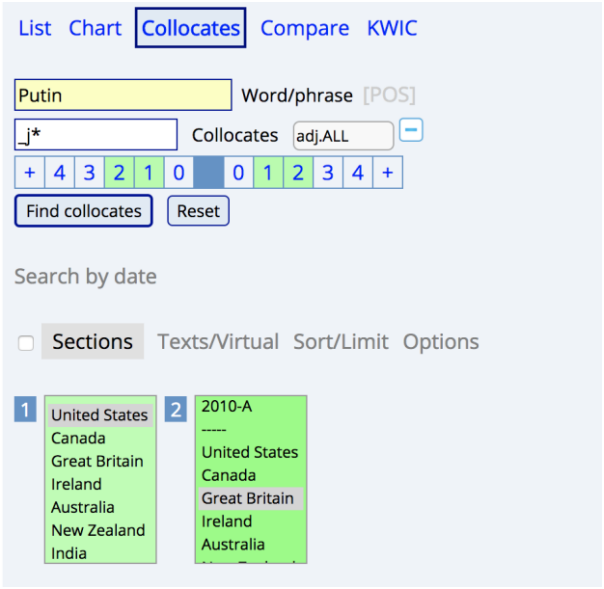
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	FLAT	44180	24967	63.2	29.3	2.2
2	APARTMENT	19075	49895	27.3	58.6	0.5

- Chart: Показывает частотность запроса по секторам (год или страна).

Пример возможного запроса приведен в разделе “Onboarding” (рис. 6b-6c). Отметим лишь, что цветами (красным, красноватым, белым, светло-зеленым, зеленым) отмечаются слова/фразы согласно их отношению SEC1:SEC2 (в нашем примере - US:AU).

- Collocates: используется для поиска слов, которые чаще всего встречаются в контексте рядом.

Пример: Найдем самые популярные прилагательные, использующиеся в контексте рядом со словом Putin в статьях американских и британских интернет-журналов (пожалуй, такая информация интересна не столько для лингвистов, сколько для политологов и социологов; впрочем, лингвист также сможет отметить негативные и положительные семантические значения):



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

SEC 1 (United States): 851,228,653 WORDS

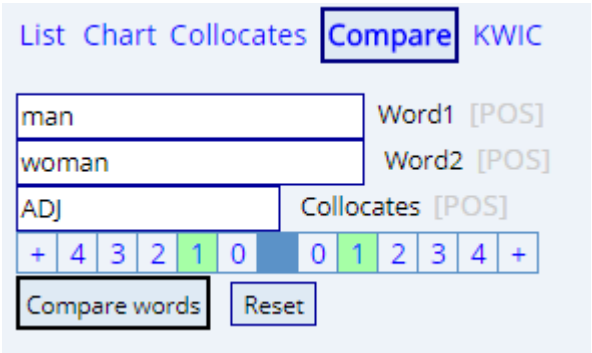
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	LONGTIME	12	1	0.0	0.0	9.8
2	FAVORITE	11	1	0.0	0.0	9.0
3	CRITICAL	10	1	0.0	0.0	8.2
4	WARY	7	1	0.0	0.0	5.7
5	FULL	12	2	0.0	0.0	4.9
6	LATE	6	1	0.0	0.0	4.9
7	GREATEST	5	1	0.0	0.0	4.1
8	HOSTILE	5	1	0.0	0.0	4.1
9	WILLING	5	1	0.0	0.0	4.1
10	SHIRTLESS	14	3	0.0	0.0	3.8
11	STRATEGIC	8	2	0.0	0.0	3.3
12	ABLE	8	2	0.0	0.0	3.3
13	APPARENT	8	2	0.0	0.0	3.3
14	WRONG	8	2	0.0	0.0	3.3
15	WORSE	4	1	0.0	0.0	3.3
16	FRESH	4	1	0.0	0.0	3.3
17	DIPLOMATIC	4	1	0.0	0.0	3.3
18	SYMPATHETIC	4	1	0.0	0.0	3.3

SEC 2 (Great Britain): 698,598,367 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	CONTROVERSIAL	11	1	0.0	0.0	13.4
2	DUE	16	2	0.0	0.0	9.7
3	HIDDEN	5	1	0.0	0.0	6.1
4	LARGE	5	1	0.0	0.0	6.1
5	SENIOR	5	1	0.0	0.0	6.1
6	WARM	5	1	0.0	0.0	6.1
7	LEFT	22	5	0.0	0.0	5.4
8	LEADING	4	1	0.0	0.0	4.9
9	IRANIAN	4	1	0.0	0.0	4.9
10	HAPPY	4	1	0.0	0.0	4.9
11	DEADLY	4	1	0.0	0.0	4.9
12	ARCTIC	4	1	0.0	0.0	4.9
13	RULING	7	2	0.0	0.0	4.3
14	SPECIAL	6	2	0.0	0.0	3.7
15	SUCCESSFUL	6	2	0.0	0.0	3.7
16	OUTSPOKEN	6	2	0.0	0.0	3.7
17	STAUNCH	3	1	0.0	0.0	3.7
18	TEMPTING	3	1	0.0	0.0	3.7

- Compare: Используется для поиска коллокаций двух слов и помогает понять их различия в значении и употреблении.

Пример: посмотрим, какие прилагательные употребляются с MAN, а какие – с WOMAN. Отсортируем по соотношению.



SORTED BY RATIO: CHANGE TO FREQUENCY

WORD 1 (W1): **MAN** (2.04)

	WORD	W1	W2	W1/W2	SCORE
1	UTD	6257	0	12,514.0	6,132.7
2	UNITED	11386	2	5,693.0	2,790.0
3	MEGA	1407	1	1,407.0	689.5
4	BETTING	534	0	1,068.0	523.4
5	WIDE	477	0	954.0	467.5
6	MACHO	724	1	724.0	354.8
7	PLAY-BY-PLAY	314	0	628.0	307.8
8	SAWAI	224	0	448.0	219.6
9	SPARE	191	0	382.0	187.2
10	MIDDLE	1477	4	369.3	181.0
11	U	324	1	324.0	158.8
12	ONE-CLUB	259	1	259.0	126.9
13	BREAFFY	129	0	258.0	126.4
14	EX-ARMY	122	0	244.0	119.6

WORD 2 (W2): **WOMAN** (0.49)

	WORD	W2	W1	W2/W1	SCORE
1	RAPED	324	1	324.0	661.1
2	SCARLET	81	0	162.0	330.6
3	CURVY	73	0	146.0	297.9
4	PREGNANT	10011	80	125.1	255.3
5	HIGHEST-RANKING	58	0	116.0	236.7
6	BIKINI-CLAD	56	0	112.0	228.5
7	PRETTY	1559	14	111.4	227.2
8	PLAN.31-YEAR-OLD	55	0	110.0	224.5
9	PRIME	541	5	108.2	220.8
10	VOLUPTUOUS	52	0	104.0	212.2
11	VEILED	200	2	100.0	204.1
12	EMPOWERED	198	2	99.0	202.0
13	MENSTRUATING	48	0	96.0	195.9
14	WONDER	1757	20	87.9	179.3

Несомненно, данный запрос также представляет бОльший интерес для социологов (лингвисту было бы интереснее сравнить коллокации, которые отражали бы различия в значении слов (оттенке) и, как следствие, их употреблении). Однако организация необходимых лингвисту запросов аналогична.

Минусы: ограничение на запрос: например, Word1 и Word2 должны содержать одинаковое количество слов, их нельзя задать в неявном виде (с помощью формул, используя «*», «?» и т.д.)

P.S: обращаем внимание на опечатки, из-за которых слова попали в список по ошибке.

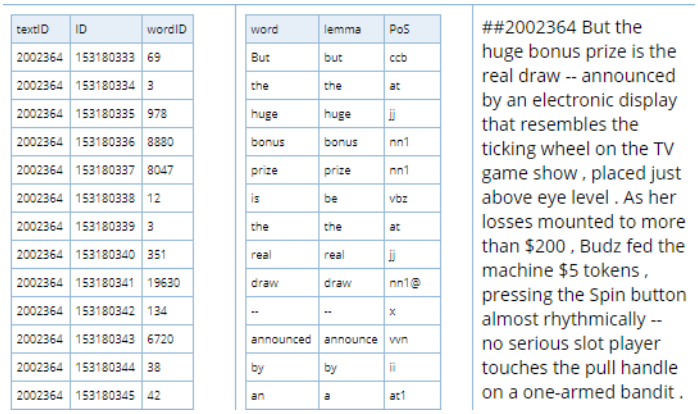
- KWIC (Keyword in Context): Позволяет увидеть контекст, в котором встречается определенное слово. При выдаче результатов поиска ближайшие к искомым формам слова в предложениях размечены по частям речи.

Также у пользователя есть возможность увидеть последние 100 добавленных вхождений для указанного в поиске запроса и найти самые популярные ключевые слова за указанный период времени (вкладки Top 100 и Keyword (by date)).

Особенности:

1. Корпус запрашивает регистрацию, однако она не требует много времени.
2. “Простой” зарегистрировавшийся пользователь имеет открытый доступ к корпусу, но в день ему позволено делать 50 запросов. Также периодически появляются объявления, призывающие приобрести подписку (индивидуальную или групповую).

3. Есть возможность скачать данные корпуса и использовать их оффлайн, однако форма, в которой предоставляются данные, неудобная (на скриншоте показан вид, в котором выдается информация):



textID	ID	wordID
2002364	153180333	69
2002364	153180334	3
2002364	153180335	978
2002364	153180336	8880
2002364	153180337	8047
2002364	153180338	12
2002364	153180339	3
2002364	153180340	351
2002364	153180341	19630
2002364	153180342	134
2002364	153180343	6720
2002364	153180344	38
2002364	153180345	42

word	lemma	PoS
But	but	ccb
the	the	at
huge	huge	jj
bonus	bonus	nn1
prize	prize	nn1
is	be	vbz
the	the	at
real	real	jj
draw	draw	nn1@
--	--	x
announced	announce	vvn
by	by	ii
an	a	at1

##2002364 But the huge bonus prize is the real draw -- announced by an electronic display that resembles the ticking wheel on the TV game show , placed just above eye level . As her losses mounted to more than $200 , Budz fed the machine $5 tokens , pressing the Spin button almost rhythmically -- no serious slot player touches the pull handle on a one-armed bandit .

Итог:

Плюсы	**Минусы**
Открытый доступ	Неудобный формат данных для работы с ними оффлайн (по крайней мере - для бесплатного пакета)
Простой и достаточно современный интерфейс	Впрочем, личный кабинет и формы создания пользовательских листов и подкорпусов оставляют желать лучшего
Несложный синтаксис поисковых запросов	Нет всплывающих окон, которые показывали бы разметку слов при наведении на них курсора мыши, отсутствие семантической и синтаксической разметки
Большое количество подсказок, инструкций, глоссарий	Отсутствие индикатора выполнения запроса, “тормознутость”
Множество функций для поиска	Не всегда возможно задать сложный запрос (с множеством указанных грамматических признаков)
Много сфер применения (не только лингвистика)	Достаточно большое количество ошибок: опечатки, случайно попавшие символы HTML-разметки, ошибочная разметка

Поработав с этим корпусом, мы можем сделать вывод, что ввиду наличия неполной разметки и часто ошибочных / «шумных» результатов поиска этот корпус не подходит для проведения глубокого лингвистического анализа, но тем не менее он фиксирует состояние современного английского языка и позволяет наблюдать за происходящими изменениями в нем изменениями. Также корпус будет полезен людям, изучающим английский язык, социологам, политологам, PR-менеджерам (отслеживать упоминания в англоязычных изданиях) и т.д.