

## **RuTenTen**

*Подготовили:*

*Ольга Веденина, группа 1, odvedenina@hse.edu.ru*

*Софья Гольдина, группа 1, smgoldina@hse.edu.ru*

*Анастасия Кузьминых, группа 1, nastia.kuzminih2011@gmail.com*

### **Общая информация**

Корпус ruTenTen относится к семейству корпусов TenTen, работающих на платформе Sketch Engine. Это корпус текстов, взятых из Интернета. Тексты представлены самые разнообразные — от комментариев на форумах до статей из Википедии. Объем ruTenTen составляет 36,946,344 документов, 1,016,579,568 предложений, 14,553,856,113 слов. Для корпусов TenTen такой объем является стандартным. Они представлены более чем для 30 языков.

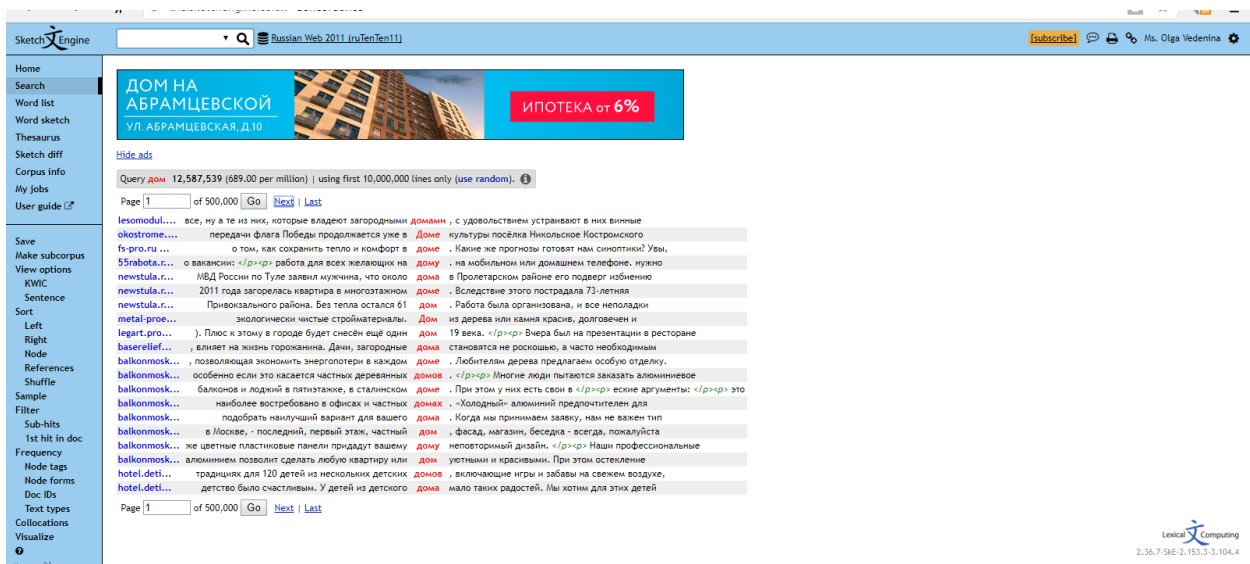
### **Дизайн**

Дизайн поисковой страницы корпуса достаточно лаконичен, цветовая гамма не раздражает глаз, не мешает восприятию текста. К минусам можно отнести наличие рекламы в бесплатной версии сайта – рекламный баннер присутствует на каждой странице поиска в верхней части страницы.

Так как корпус RuTenTen создан на платформе Sketch Engine, пользователям, работавшим с корпусом Aranea, не составит никакого труда разобраться в устройстве этого корпуса.

Сайт корпуса интуитивен, благодаря чему пользователю, впервые пользующемуся этим ресурсом, будет просто разобраться в навигации по сайту. Как и на многих других сайтах, меню корпуса расположено в левой части экрана. В правой верхней части экрана находятся значки, позволяющие оформить подписку на сайт, оставить фидбэк, распечатать содержимое страницы, копировать сжатую ссылку на страницу, а также меню "Настройки". Их назначение сразу понятно, однако увидеть их можно не сразу из-за маленького размера.

Еще один недостаток в дизайне сайта – невозможность увеличить или уменьшить количество примеров в выдаче. Например, зададим поиск по лемме "дом". В выдаче мы получим 20 примеров на страницу. Чтобы посмотреть дальше, необходимо листать.



## Onboarding (ресурс глазами новичка)

Корпус достаточно легко найти в интернете, хотя может возникнуть небольшая путаница из-за того, что открывается не сам корпус, а сайт платформы Sketch Engine, и нужно понять, как перейти к поисковой странице. Чтобы воспользоваться корпусами TenTen, необходима регистрация. Бесплатное использование доступно в течение 30 дней, после этого придется оформлять подписку. Сайт корпуса интуитивен, благодаря чему пользователю, впервые пользующемуся этим ресурсом, будет просто сориентироваться в навигации по сайту. Что касается поиска, в нем очень легко разобраться практическим путем. Кроме того, ресурс предоставляет пользователям огромное количество самых разнообразных инструкций и подсказок, подробнее о которых будет сказано ниже.

## Помощь

Для неопытных пользователей корпусов на сайте есть целый обучающий раздел по использованию корпуса. Найти его можно несколькими способами:

1. В меню в разделе "User guide". Здесь есть целый набор обучающих инструкций, разбитый по разделам – основные функции, видео-гайды и т.д.
2. В поиске по корпусу есть кнопка – знак вопроса. Наведя на нее мышкой, мы увидим всплывающую надпись "How to search in corpus?" Щелкнув по ней, сайт отправит нас на страницу, где содержится вся обучающая информация по поиску в корпусе. Кроме этого имеется справочник по регулярным выражениям и CQL поиску.

Сторонних гайдов и тьюториалов по использованию этого корпуса найти не удалось, однако, на наш взгляд, информации, которая содержится на сайте корпуса, достаточно. Единственный ее минус – гайды составлены на английском языке. Поэтому русскоязычных гайдов для этого корпуса нет.

## Продвинутый функционал

Корпус обладает двумя уровнями разметки — по леммам и по частям речи. Кроме того, тексты проходят чистку в `juText`, где из них убирают ненужные части, например, рекламу и ссылки. Затем их токенизируют и избавляют от повторяющихся фрагментов.

Помимо стандартных функций (точного поиска, поиска по лемме, словоформе, букве и фразе) есть также возможность писать запросы, пользуясь языком CQL (Corpus Query Language). Кроме этого, можно отфильтровать контекст интересующего вас слова по лемме и части речи и задавать домен сайтов, по которым вы хотите провести поиск. Полученную выдачу можно скачать, отсортировать разными способами, представить в виде графика и составить случайную выборку. Также можно ознакомиться со списком коллокаций для заданной леммы.

Sketch Engine Russian Web 2011 (ruTenTen11)

Query **пить, ru** 1,430,254 > Positive filter (excluding KWIC) **часто** 1,922 > Positive filter (excluding KWIC) **N.\*** 660 (0.04 per million)

Page 1 of 33 Go Next Last

Indigal.ru... пишу употребляете недостаточно; **</p><p> часто пьёте пиво**; **</p><p>** курите. **</p><p>** Основной причиной развития  
brimz.ru % эффективностью. Когда мы не забываем **часто пить воду**, мы даем возможность нашему метаболизму  
pragaonlin... к продаже в других странах. **</p><p>** Чехи много и **часто пьют кофе**. В многочисленных кофейнях подают кофе  
jhana.ru писатель Северной династии Сонг **часто пили чай** и пели стихи, некоторые из которых  
darkdiary... мания всем находить занятия и задания. Мы **часто пили чай** в столовой все вместе, и в такие моменты мне  
kuppo.ru ивский князь Владимир Святославович **часто пил напиток**, называемый «кава». Правда сегодня не  
vrednoli.ru... **</p><p>** Если задаться вопросом, вредно ли **часто пить чай**, то ответ может не порадовать (обычно это  
vrednoli.ru... этого напитка до 2-3-х чашек. Вредно ли **часто пить чай**? В больших количествах вредно, но если Вы  
www2.newsa... быть даже полезным и у французоз, которые **часто пьют вино** гораздо ниже вероятность инфаркта и  
rusbereza... слезы счастья и радости! Желаю Вам очень **часто пить напиток** радости и счастья! Пусть все Ваши  
detka-konf... не нужно снимать даже в воде. Пусть малыш **часто пьет воду**, чай, компот или морс, за исключением  
city-cater... По, любитель абсента и вообще алкоголик, **часто пил смесь** абсента и бренди со своим издателем  
samaraham... хмзов. Позже, с институтскими **часто пили пиво** на ул.Арцыбушевской, откуда были хорошо  
bukharapit... Илья. У папа было много денег, они с Илей **часто пили спирт**. Поп пил только спирт. Максим пошел к  
kabluchok... аперитивом, а именно в этом качестве мы **часто пьем шампанское** на Новый год. В следующем затем  
tourhot.ru... случаях крепости 52о. Сами словари не так **часто пьют сливовицу**, предпочитая ей Боровичку - джин  
drugs.com... без перерыва (при этом непременно нужно **часто пить воду** или сок!). Чувство единства и разделения  
lab.tekora... думаешь, что маркетологи из Альфа-Банка **часто пьют кофе** в Starbucks и заходят на сайт этой  
znaha.ru заявку на info@frio.ru **</p><p>** Женщины, которые **часто пьют пиво**, рискуют быть «пожалованными»  
moepivo.ru... он любит, куда ходит, как отдыхает, как **часто пьет пиво**, каков его уровень дохода, какую сумму он

Page 1 of 33 Go Next Last

Вот пример обычного поиска в ruTenTen (в данном случае мы хотим найти словосочетания вида «часто + пить + сущ») на сайтах с доменом .ru).

Стоит также обратить внимание на функцию word sketch. Word sketch позволяет найти список слов, с которыми интересующая нас лемма чаще всего употребляется. Слова распределены по группам в зависимости от роли, которую они исполняют по отношению к искомой лемме в предложении. Их можно отсортировать по частотности внутри каждой группы или вывести общим списком. Вот как выглядит выдача для слова «пить»:

Sketch diff	Hide ads
Corpus info	
My Jobs	
User guide	
Save	
Change options	
Cluster	
Sort by freq	
Hide gramrels	
More data	
Less data	
Sketch grammar	
Translate	
- French	

<b>ПИТЬ</b> (verb) Alternative PoS: noun (67,481) adjective (1,715)		Russian Web 2011 (ruTenTen11) freq = 1,363,768 (74.60 per million)	
subject	8.48	object4	25.93
пиво +	2,785 6.58	чай +	60,012 10.22
Кофе +	346 6.35	пить чай	
Кофе пьют		пиво +	30,310 9.70
настой +	782 6.25	пить пиво	
глинтвейн +	266 6.07	кофе +	27,326 9.59
пили глинтвейн		прописали пить	
вино +	2,260 5.77	водка +	13,405 9.11
вегетарианец +	266 5.66	пить водку	
" пьют ли вегетарианцы		таблетка +	8,872 8.53
самогон +	220 5.60	пить таблетки	
чай +	490 5.57	пить	17,457 8.50
verb_post_inf	12.48	verb_post_inf	12.48
бросить +	5,429 8.24	бросил пить	
бросать +	7,159 7.88	бросить пить	
прописать +	1,139 7.34	прописали пить	
рекомендоваться +	4,137 7.20	рекомендуется пить	
перестать +	3,829 6.79	перестал пить	
перестать пить		перестал пить	
post_prep	10.98	post_prep	10.98
вместо +	1,217 4.93	вместо	
пить вместо		сверх	89 3.58
жадно +	1,947 8.49	пить сверх меры	
много +	1,782 7.77	перед +	2,313 3.53
маленький +	1,839 7.69	пить перед	
полезно +	1,291 7.50	из +	18,107 3.38
полезно пить		пить из	
чай +	639 7.31	по +	30,334 3.21
чай пить		Пить по	
adv_modifier	9.53	adv_modifier	9.53
жадно +	1,947 8.49	жадно +	1,947 8.49
много +	1,782 7.77	много +	1,782 7.77
маленький +	1,839 7.69	маленький +	1,839 7.69
полезно +	1,291 7.50	полезно +	1,291 7.50
полезно пить		чай +	639 7.31
чай +	639 7.31	чай +	639 7.31

Следующая важная функция — word list. Она формирует частотный список слов по всему корпусу. Ей можно задавать некоторые параметры, например, часть речи или использование n-грамм.

Thesaurus — функция, позволяющая найти слова, синонимичные заданному или схожие с ним по категории. Они выдаются списком или словесным облаком.

Последняя функция, о которой хотелось бы упомянуть — sketch diff. С ее помощью можно сравнить два синонимичных слова по коллокациям, в составе которых они употребляются. Вот пример для слов «красивый» и «прекрасный»:

красивый/прекрасный (adjective) Alternative PoS: noun (freq: 6,627) verb (freq: 2,856)  
Russian Web 2011 (ruTenTen11) freqs = 2,548,521 | 2,164,170

красивый	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	прекрасный
----------	-----	-----	-----	---	------	------	------	------------

и/или	994,762	413,444	0.39	0.19
удобный	21,829	647	7.9	3.2
прочный	5,549	170	6.8	2.3
модный	6,206	191	6.9	2.5
качественный	11,072	542	6.9	2.8
стильный	10,564	380	7.9	3.9
практичный	4,082	144	6.8	2.9
здоровый	21,334	1,223	8.1	4.3
ровный	6,499	325	7.2	3.6
привлекательный	5,766	317	6.9	3.3
оригинальный	8,615	615	7.0	3.7
стройный	13,809	710	8.5	5.2
умный	22,068	1,557	8.7	5.5
уютный	11,339	789	7.9	4.7
элегантный	4,720	295	7.0	3.9
яркий	18,119	1,872	7.7	4.7
вкусный	4,337	353	6.6	3.7
интересный	13,555	1,549	6.8	3.9
ухаженный	8,719	559	8.0	5.2
молодой	36,908	4,628	7.9	5.1
необычный	6,810	819	7.0	4.5
богатый	7,344	1,184	6.8	4.6
изысканный	4,790	564	7.0	4.8
женский	8,380	1,657	6.6	4.6
чистый	11,458	4,142	6.7	5.5
нежный	4,755	2,018	6.6	6.0

adv_modifier	368,495	66,353	0.14	0.03
очень	175,054	1,902	8.7	2.2
потрясающе	7,027	183	9.1	5.9
необычайно	7,609	322	9.0	5.9
безумно	5,109	285	8.5	5.8
невероятно	9,795	676	9.2	6.5
намного	2,537	285	6.3	3.7
необыкновенный	8,120	493	9.2	6.8
внешне	972	73	6.1	3.8
исключительно	1,685	213	6.3	4.0
удивительно	11,281	1,041	9.5	7.4
поразительно	1,190	77	6.6	4.8
особенно	6,620	1,621	6.7	4.9
чертовски	899	63	6.2	4.4
удивительный	1,785	106	7.3	5.5
изумительно	2,615	205	7.8	6.4
по-настоящему	1,637	364	6.7	5.8
действительно	5,613	3,165	6.4	5.8
настолько	3,282	1,744	6.5	6.0
мне	4,712	2,893	6.5	6.0
фантастически	2,123	374	7.5	7.2
сказочно	3,965	879	8.4	8.5
ослепительно	1,414	427	6.9	7.3
эстетически	747	223	6.0	6.5
по-своему	1,242	986	6.6	8.1
божественно	849	650	6.2	8.2

subj_быть	4,057	5,351	0.00	0.00
петлюра	44	0	8.4	--
Моргот	12	0	6.1	--
Эрнесто	15	0	5.8	--
почерк	11	0	1.9	--
зрелище	19	0	1.4	--
закат	15	0	1.4	--
жених	17	0	1.1	--
юноша	16	0	0.1	--
девушка	85	0	0.1	--
замысел	0	13	--	0.1
невеста	0	14	--	0.1
ночь	0	89	--	0.4
самочувствие	0	11	--	0.5
завтрак	0	23	--	0.7
вечер	0	98	--	0.7
утро	0	75	--	0.8
девица	0	10	--	1.3
видимость	0	24	--	1.6
настроение	0	132	--	1.6
принцесса	0	32	--	2.3
слышимость	0	10	--	4.2
погода	0	600	--	4.2
Анакин	0	12	--	5.6

красный»:

## Закключение

Корпус ruTenTen, безусловно, обладает огромной ценностью для лингвистических исследований. Однако он далеко не идеален, и в нем есть существенные недостатки, которые затрудняют работу. Во-первых, разметка часто выполнена неправильно. Мы убедились в этом на примере коллокаций для слова «пить», где много ошибок при выделении подлежащего и дополнения. Например, во фразе «чай пьет» «чай» с большой долей вероятности будет обозначен как подлежащее. Во-вторых, в текстах сохранена авторская орфография, а так как они взяты из интернета, то, естественно, в них встречается много ошибок и опечаток. Это создает множество проблем, в частности, синтаксическую омонимию, результатом которой является погрешность в выборке. И наконец, как нам кажется, корпусу не хватает тематической классификации текстов, как,

например, в НКРЯ. Часто для исследования необходим какой-то определенный тип текстов, а в ruTenTen нет подходящих подкорпусов для таких запросов.