корпорация

российский учебник

rosuchebnik.ru

корпорация Российский учебник

Национальный корпус русского языка: как использовать сервис в школе?

Михаил Иванович Шаповалов

Национальный корпус русского языка

http://www.ruscorpora.ru

Национальный корпус русского языка

Подкорпуса

основной

– корпус

– биграммы

– триграммы

– 4-граммы

– 5-граммы

синтаксический

газетный

параллельный

обучающий

диалектный

поэтический

устный

акцентологический

мультимедийный

мультипарк

исторический

Распределение текстов по подкорпусам

Подкорпус	Число текстов	Число предложений	Число словоупотреблений
Основной корпус	76 882	17 574 752	209 198 275
- в том числе со снятой омонимией	2 147	516 852	5 944 188
Газетный корпус	181 175	8 553 495	113 292 003
Диалектный корпус	197	20 273	194 283
Обучающий корпус	229	65 666	664 751
Параллельный корпус	370	1 609 609	24 022 437
Поэтический корпус	41 448	638 861	6 738 474
Устный корпус	3 034	1 604 626	10 122 579
Мультимедийный корпус	31 741	148 619	648 576
Всего:	335 076	30 215 901	364 881 378

Параллельный корпус

параллельный

– английский

– армянский

– белорусский

– болгарский

– бурятский

– испанский

– итальянский

– китайский

– латышский

– литовский

– немецкий

– русская классика
в немецких переводах

– польский

– украинский

– французский

– шведский

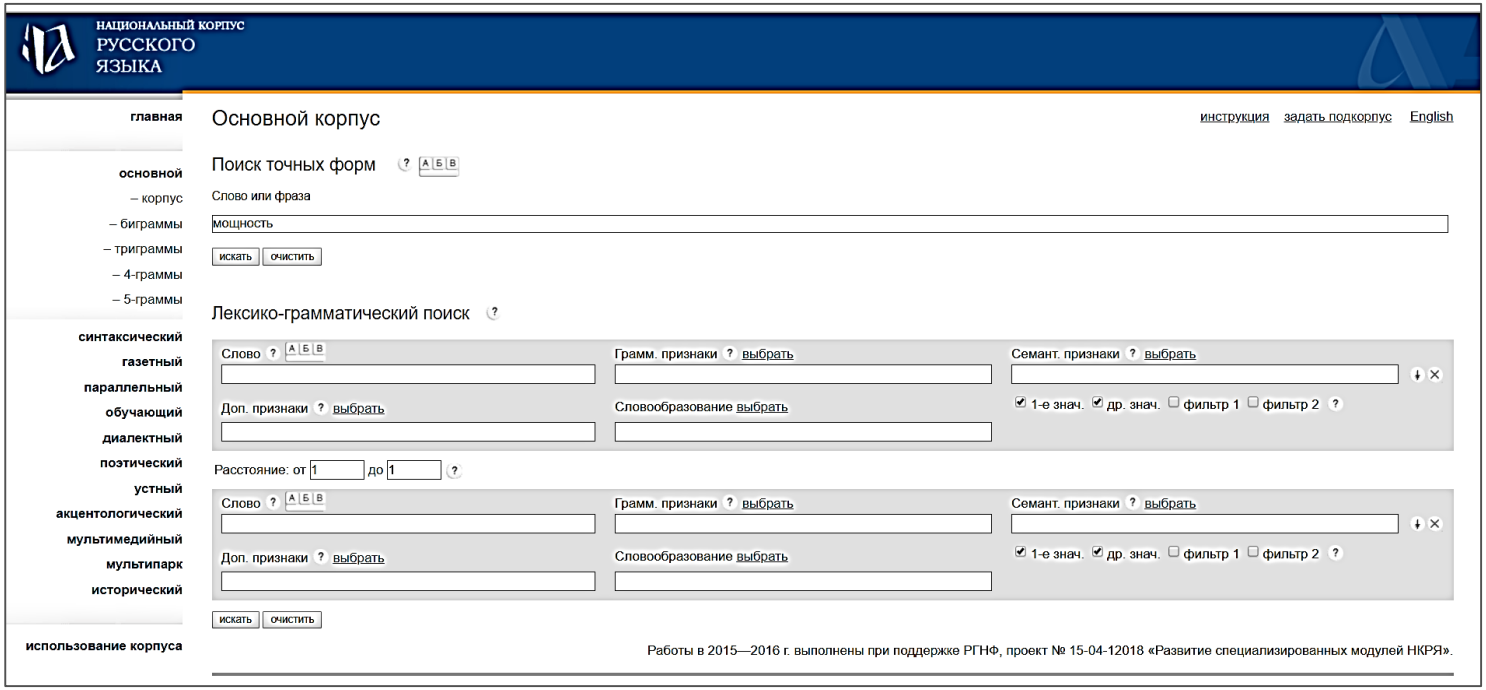
– эстонский

– многоязычный

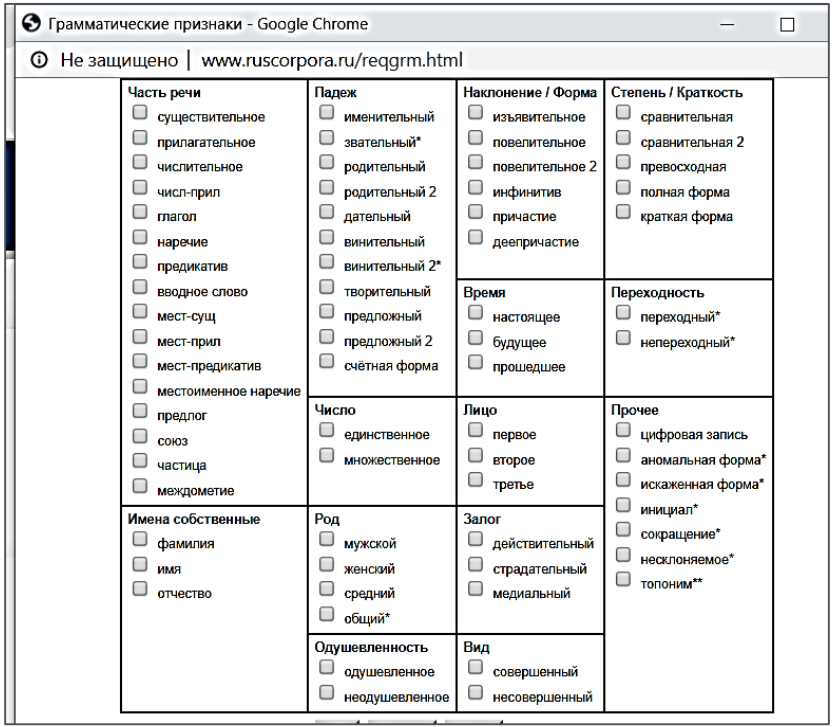
Распределение текстов по видам

Вид текста	Число текстов	Число предложений	Число словоупотреблений	% словоупотреблений
Художественные письменные тексты	6 308	9 164 402	92 058 363	44.0%
Нехудожественные письменные тексты	70 574	8 410 350	117 139 912	56.0%
Всего:	76 882	17 574 752	209 198 275	100

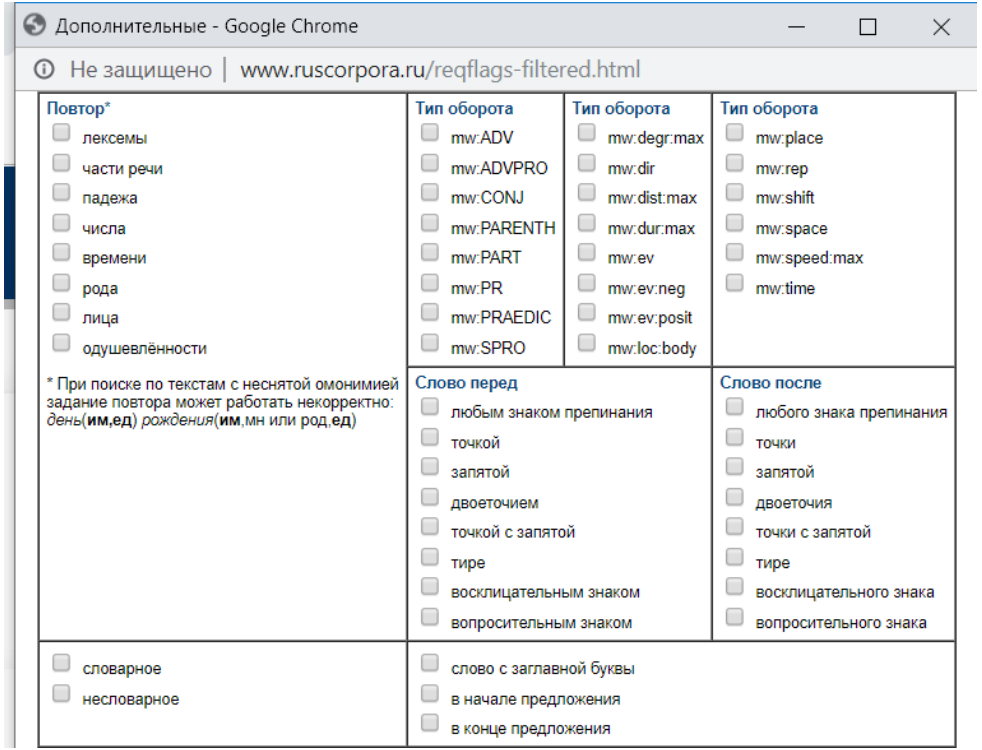
Поиск точных форм



Грамматические признаки



Дополнительные признаки



Использование корпуса

Данные корпусов могут быть использованы для построения и уточнения грамматик и в целях обучения языку.

Н.Р. Добрушина отмечает, что «в типичном учебнике русского языка упражнения на 60% состоят из примеров XIX века; 30 % приходится на литературу XX века, а еще 10% составляют примеры, сконструированные автором.

Но для общего развития школьников нужно включать в учебный материал образцы тех текстов, которые имеют отношение к современной жизни, а не только к литературе»

Использование корпуса

На основе Корпуса может быть организована самостоятельная работа учащихся.

Педагогический эффект такой работы очевиден: «…одно дело выполнить «пассивную» работу: определить частеречную принадлежность слова в тексте, подобранном преподавателем и совсем другое - «активная» работа: самому отыскать в Корпусе случаи употребления заданного слова и выбрать из них хотя бы по одному на каждую возможную часть речи.

Задания «активного типа» требуют от учащихся гораздо большей работы мысли, оказываются более интересными и полезными»

Примеры заданий

1. Среди существительных путь, степень, ступень, ставень найдите такие, которые изменили свою родовую принадлежность. Когда это произошло?

2. Найдите в НКРЯ примеры ошибочных постановок запятых при невводных словах, омонимичных вводным (наконец, однако, действительно и т. д.).

3. С помощью Национального корпуса русского языка выберите по 20 оценочных прилагательных с положительной и отрицательной коннотацией.

Примеры заданий

1. Проверить по Национальному корпусу русского языка следующую гипотезу:

Гипотеза: Когда глагол согласуется с количественной группой (типа "два человека", "много людей", "несколько чашек", "три автобуса"), то выбирается форма множественного числа, если существительное в количественной группе – одушевленное (пришли два человека), и форма среднего рода единственного числа, если существительное – неодушевленное (упало несколько чашек).

Примеры заданий

1. Когда появилось слово Фиолетовый. Когда появилось слово Сиреневый в значении цвета.

2. Дайте запрос: «белый + существительное». Разграничить свободные словосочетания и фразеологизмы.

Использование данных Корпуса при изучении лексики

Корпус существенно облегчает задачи преподавателя при составлении заданий по курсу лексики. Преподаватели знают, что многие группы слов находятся у школьников в пассивном словарном запасе, а иногда и вовсе оказываются незнакомыми и непонятными для них.

Это относится, например, к разговорной, стилистически окрашенной лексике: ротозей, гоготать, обмишуриться и др.

Это редко встречающиеся слова, и подобрать примеры их употребления из текстов - достаточно трудная задача. Корпус в течение нескольких секунд выдает все имеющиеся в нем предложения с этими словами

Примеры из Корпуса

1. Полицейский на съемочной площадке призван обеспечивать порядок, оберегая нас от **ротозеев**, но тот же самый полицейский строго следит за тем, чтобы киногруппа не нарушала распоряжения городских властей (Родион Нахапетов. Влюбленный (1998)).

2. Проверяющие заволновались, поняли, что **обмишурились**, попытались объяснить мне, что находятся здесь по решению райкома партии, что обязаны зафиксировать тех, кто опоздал или вообще не явился на работу (Александр Яковлев. Омут памяти (2001)).

3. Они одинаково бесились, у всех мелькали одни и те же словечки, шуточки, они одинаково хохотали, вернее **гоготали** (Даниил Гранин. Зубр (1987)).

Паронимы

При изучении темы Паронимы, используя возможности Корпуса, можно составить эффективные упражнения.

При помощи найденных в корпусе глаголов-паронимов представить и предоставить в разных формах можно составить упражнение на их различение. Например:

1. Российский Минобраз _____ (предоставил) регионам право самим решать, при какой температуре воздуха дети могут не ходить в школу (Когда градус мешает учебе (2003) // «Вечерняя Казань», 2003.01.11).

2. Когда зал наполнился приглашенными, ведущий открыл встречу и _____ (предоставил) слово заместителю губернатора области (Александр Шептуха. Большой строительный праздник (2003) // «Пермский строитель», 2003.09.10).

3. Руководитель департамента _____ (представил) почти ста участникам круглого стола строительные возможности и перспективы Краснодарского края (Ольга Гордеева, Игорь Моржаретто. После спячки (2004) // «За рулем», 2004.04.15).

Паронимы

4. Музей в пятый раз безвозмездно _____ (предоставил) помещение для торжественной церемонии (С миру по нитке - это и есть сейчас (2002) // «Культура», 2002.03.25).

5. Суду были также _____ (представлены) документы, отражающие все этапы борьбы жильцов дома с проведением незаконных и опасных ремонтных работ (Ольга Мозговая. Главное - чтобы костюмчик «ушел»! (2002) // «Вечерняя Москва», 2002.08.08).

6. Сегодня управляющий должен _____ (представить) список всех дворовых (А. А. Потехин. В мутной воде (1871)).

Таким образом, можно сделать вывод, что предоставить - это «отдать в распоряжение; дать какое-нибудь право, возможность», а представить - «доставить, предъявить, сообщить».

Пример использования «даровитый – талантливый»

Почти нет примеров употребления прилагательного даровитый с существительными, обозначающими тексты. Во всем Корпусе всего два таких примера

А вот слово талантливый свободно сочетается как с названиями лица, так и с названиями текстов

Семантическая классификация метафор

Используя данные Корпуса, несложно подобрать богатый и разнообразный материал по теме «Семантическая классификация метафор» на основе современных публицистических текстов.

Можно предложить учащимся выделить среди данных примеров различные виды метафор (медицинские, спортивные, театральные, военные, финансовые, метеорологические):

1. Ключевая причина высокой хронической безработицы в Европе - структурные проблемы, чрезмерная зарегулированность экономики и, в частности, рынка труда (Александр Кокшаров. Одной ногой в рецессии (2004) // «Эксперт», 2004.12.13).

2. События в Косово и в Тибете сразу же вошли в арсенал средств, используемых в глобальной геополитической борьбе» (Сергей Маркедонов. Кавказские приоритеты внешней политики Казахстана // «Неприкосновенный запас», 2009).

Омонимия. Мощность

А. А. Антонов. Минералогия родингитов Баженовского гипербазитового массива (2003)

Мощность жил варьирует от нескольких до 40 см. Хромитит имеет полиэдрическую структуру, средний размер зёрен около 3-4 мм.

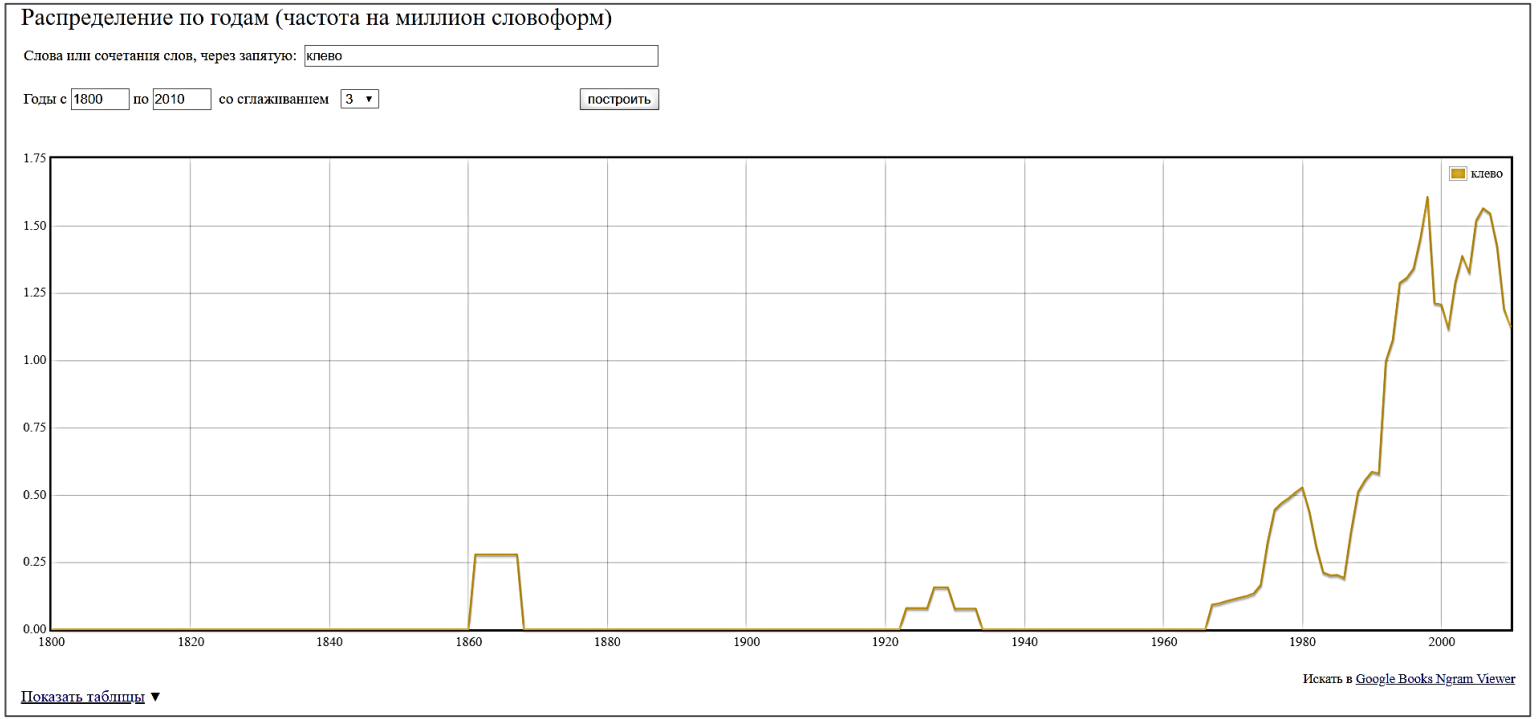
Владимир Лукашик, Елена Иванова. Сборник задач по физике. 7-9 кл. (2003)

При этом **мощность**, развиваемая двигателями, и число оборотов винта не изменились.

Игорь Лалаянц. Детектор лжи на молекулярном уровне? Завтра, завтра… послезавтра! // «Знание -- сила», 2003

Ко времени его появления резко повысилась **мощность** компьютеров

График

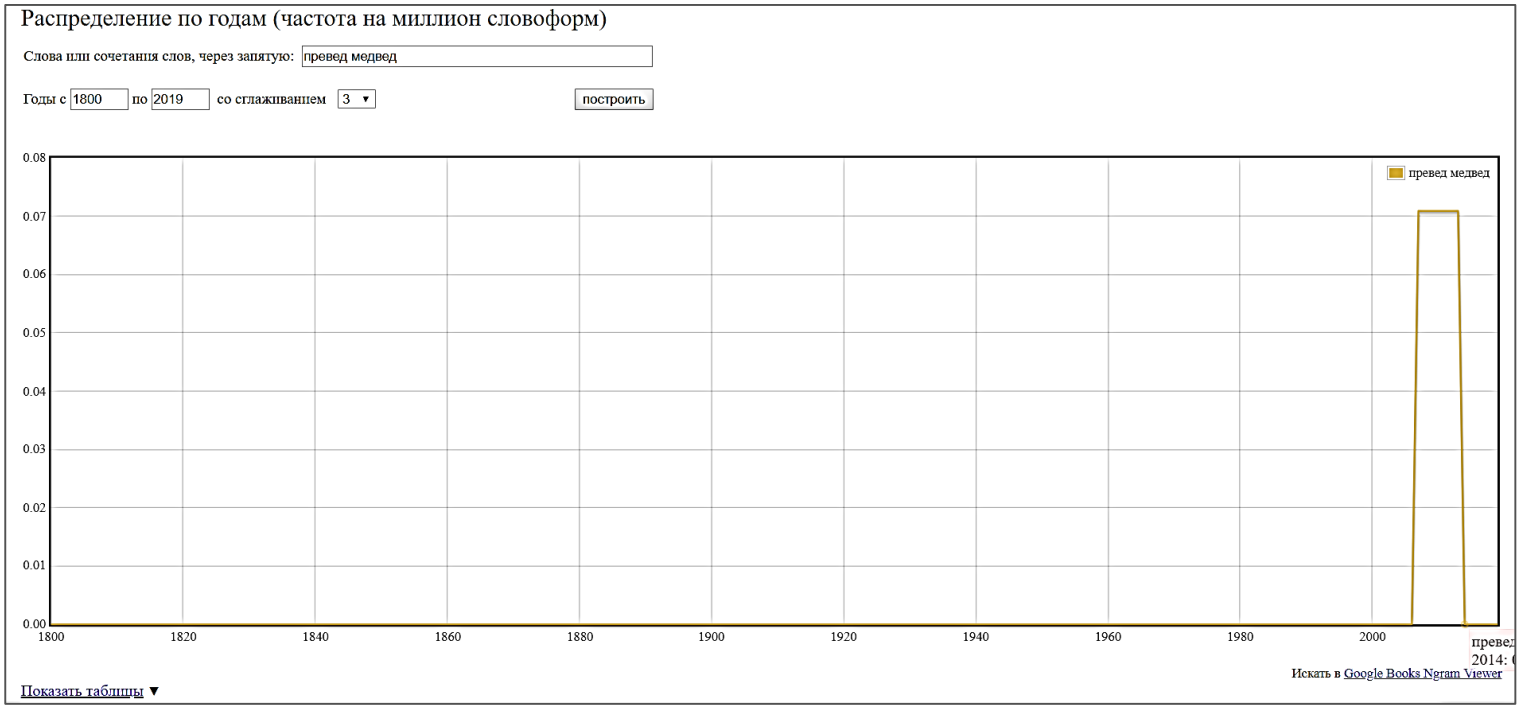


Примеры пословиц на языке офень

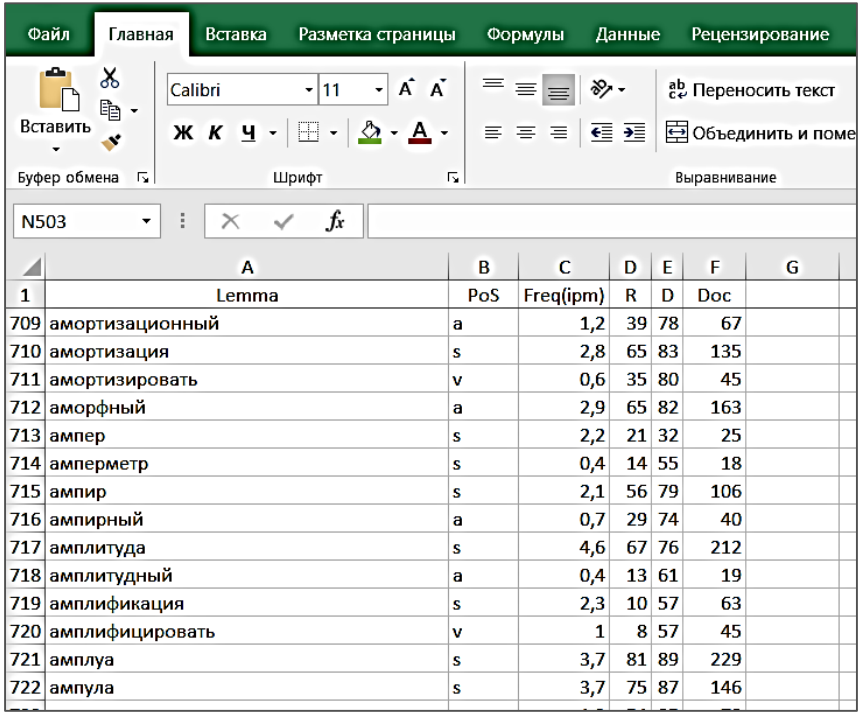
1. Век живи, век учись – дураком помрешь. – Пехаль киндриков куравь, пехаль киндриков лузнись – смуряком отемнеешь.
2. Кто не работает – тот не ест. – Кчон не мастырит, тот не бряет.
3. Без труда не выловишь и рыбку из пруда. – Без мастыры не подъюхлишь и псалугу из дрябана.
4. Прошел огонь, воду и медные трубы. – Прохлил дрябу, дулик и фильные фошницы.
5. Ни кола, ни двора. – Ни брута, ни рыма.
6. Муж и жена – одна сатана. – Муслень и елтона – ионый кульмас.

- Источник: Кто такие офени
-

Олбанский язык. Превед медвед



Коэффициент Жуйана D



	A	B	C	D	E	F	G
1	Lemma	PoS	Freq(ipm)	R	D	Doc	
709	амортизационный	a	1,2	39	78	67	
710	амортизация	s	2,8	65	83	135	
711	амортизировать	v	0,6	35	80	45	
712	аморфный	a	2,9	65	82	163	
713	ампер	s	2,2	21	32	25	
714	амперметр	s	0,4	14	55	18	
715	ампир	s	2,1	56	79	106	
716	ампирный	a	0,7	29	74	40	
717	амплитуда	s	4,6	67	76	212	
718	амплитудный	a	0,4	13	61	19	
719	амплификация	s	2,3	10	57	63	
720	амплифицировать	v	1	8	57	45	
721	амплуа	s	3,7	81	89	229	
722	ампула	s	3,7	75	87	146	

Пример работы с частотным словарем

1. Провести лемматизацию трех текстов, используя интернет-сервис Advego http://advego.ru/text/seo/
2. Провести лемматизацию и найти частоты слов трех текстов, используя интернет-сервис К50 https://snacks.k50.ru/lemmatizeIndex.php
3. Определить частоты встречаемости слов для трех текстов с использованием частотного словаря Корпуса русского языка.
4. Сделать вывод об авторстве трех текстов, путем сравнения приведенных частот встречаемости слов.

Анализ частот семантического ядра

Каким должно быть соотношением частей речи в тексте обсуждалось еще в Древнем Риме.

Считалось, что "золотым сечением" является пропорция, при которой объединяются один глагол с двумя существительными и двумя прилагательными. Предполагалось, что такую фразу отличают легкость понимания, ясность, четкость и прозрачность.

Можно легко сравнить частоту встречаемости слов в тексте с частотой по Национальному корпусу русского языка.

Анализ частот семантического ядра

Например, для эссе «Теория травмы — первое знакомство» (1 курс университета) количество существительных 48%.

В Национальном корпусе русского языка существительные составляют 28,5%. Т.е. обычный процент существенно превышен (почти в два раза).

Легко видеть (последний столбец), что частоты отдельных слов превышают средние по Корпусу от 10 до 137 раз. В теории SEO-оптимизации такое явление называют «переспам» и оценивают крайне негативно.

Анализ семантического ядра

	Количество	Частота	Пересчет на миллион	По частотному словарю Корпуса русского языка	Соотношение (превышение частоты)
травма	22	2,44	2702,10	19,6	137,86
событие	6	0,66	730,90	206,5	3,54
боль	5	0,55	609,08	113,3	5,38
помощь	4	0,44	487,26	300,1	1,62
страдание	4	0,44	487,26	39,8	12,24
призыв	3	0,33	365,45	33,9	10,78
психика	3	0,33	365,45	20,2	18,09
смерть	3	0,33	365,45	284,1	1,29
грубый	2	0,22	243,63	42,6	5,72
жесткий	2	0,22	243,63	73,6	3,31
мучение	2	0,22	243,63	10,1	24,12

Информационная насыщенность

Сервис семантического анализа Advego выдает для текста такие параметры, как количество уникальных слов и общее количество слов.

Казалось бы, их отношение можно использовать для оценки информационной насыщенности текста (чем больше отношение, тем выше информационная насыщенность).

Однако, правильнее учитывать коэффициент Жуйана, значение которого зависит от того, является ли слово термином. Именно термины передают основной смысл текста, представляя собой информацию в «сжатом» виде. Коэффициент Жуйана является лучшим из известных в настоящее время способов измерить, насколько общеупотребительным является слово, или, напротив, насколько оно специфично для отдельных предметных областей.

Оценка активного словарного запаса

Важной характеристикой следует считать размер активного словарного запаса учащегося. Как и в случае, когда мы рассчитывали коэффициент информационной насыщенности, недостаточно учитывать только частоту слов, используемых автором в тексте - это значение мы получаем из частотного словаря Национального корпуса русского языка (НКРЯ). Следует учитывать и частоту встречаемости слов в разных разделах частотного словаря (значение коэффициента Жуйана НКРЯ), т.е. частоту встречаемости в текстах разных жанров.

Тексты в корпусе упорядочиваются по функциональным стилям, поэтому «тексты одного жанра (например, научные статьи) аккумулируются в пределах небольшого числа сегментов».

Так как когда мы говорим о величине словарного запаса мы, в первую очередь, имеем в виду слова общей лексики, оценка активного словарного запаса тем выше, чем более редкие слова (общей лексики) использует автор, а коэффициент Жуйана D позволяет снизить влияние терминов.