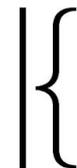


# Как корпусная лингвистика помогает изучать эволюцию языка

---

Дмитрий Морозов  
м.н.с. ЛабПЦТ ММЦ НГУ | технический директор НКРЯ

 Russian  
NATIONAL  
CORPUS



**LABADT**

# План на сегодня

1. Что такое корпусная лингвистика?
2. Что такое Национальный корпус русского языка?
3. НКРЯ и ML
4. НКРЯ и эволюция русского языка

Что такое корпусная лингвистика?

# Что такое корпус текстов?

Собрание текстов в электронной форме с широкими возможностями поиска по разным параметрам.

Обычно тексты для корпуса выбираются не случайным образом, а исходя из его назначения.

# Почему корпус — не библиотека

Электронная библиотека — для поиска литературы по метаатрибутам и её чтения.

Корпус текстов — для поиска и анализа примеров употребления слов и словосочетаний с богатой системой внутритекстового поиска.

# Почему не интернет-поиск?

- богатые возможности синтаксиса языка запросов
- специальная мета- и внутритекстовая разметка
- поиск, просмотр и подсчёт всех примеров

## Какие бывают корпуса?

Первый корпус — Брауновский (1963 г.). Содержит 500 текстов по 2000 слов каждый. Тексты выбраны из книг, газет, журналов, издававшихся в период с 1961 по 1963 год в США. Корпус призван отражать многообразие английского (американского) языка.

# Какие бывают корпуса?

- Британский Национальный Корпус
- Чешский Национальный Корпус
- SketchEngine-корпуса
- Reverso Context
- Google Books Ngram Viewer
- ...



# Зачем нужны корпуса?

Фактически, любой корпус является виртуальным носителем языка.

Задачи, решаемые при помощи корпусов:

- изучение языка лингвистами
- проверка лингвистических гипотез при редактировании текста
- преподавание языка как родного и как иностранного
- обучение переводчиков
- обучение языковым моделям, в том числе, обучение моделей для перевода
- ...

Что такое НКРЯ?

# Машинный фонд русского языка

Проект создания корпуса русскоязычных текстов и разработки средств автоматического лингвистического анализа.

Начат в 1985 году по инициативе академика А. П. Ершова (1931-1988).

По ряду причин успеха не получилось.



# Машинный фонд русского языка

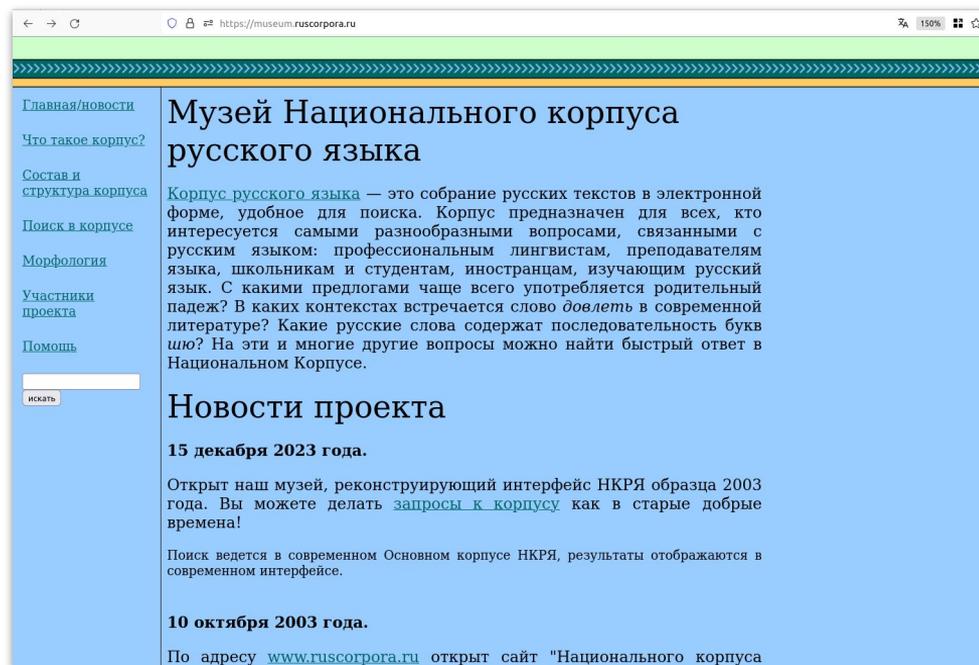
*Сейчас, в 2005 г. мы должны признать, что данное научное направление (информатизация русистики) оказалось нежизнеспособным в современных организационно-финансовых условиях и постановка задачи создания Машинного фонда русского языка на ближайшую перспективу должна быть еще более сужена до двух-трех частных задач.*

В. М. Андрющенко

# Национальный корпус русского языка

Корпус с начала 2000-х собирался и размечался командой лингвистов, а программировали корпусный движок специалисты из Яндекса — И. В. Сегалович и В. А. Титов.

За многие годы ключевую роль в развитии Корпуса сыграли ИРЯ РАН, ИППИ РАН и Яндекс.



# Юбилей Корпуса

29 апреля 2024 года НКРЯ  
исполнилось 20 лет!

https://ruscorpora.ru

Национальный корпус русского языка — представительная коллекция текстов *на русском языке* общим объемом *более 2 млрд слов*, оснащенная лингвистической *разметкой* и инструментами поиска

ИРЯ Яндекс

[Подробнее о Корпусе](#)

Введите слово или фразу  [Обзор возможностей](#)

### Поиск по корпусам

<a href="#">Основной</a> (374 млн)	<a href="#">Устный</a> (13 млн)	<a href="#">Параллельные</a> <sup>28</sup> (179 млн)	<a href="#">Поэтический</a> (13 млн)
<a href="#">Газетные</a> <sup>2</sup> (850 млн)	<a href="#">Акцентологический</a> (135 млн)	<a href="#">Диалектный</a> (599 тыс)	<a href="#">Русская классика</a> $\beta$ (18 млн)
<a href="#">Синтаксический</a> (1,6 млн)	<a href="#">Мультимедийный</a> (5,8 млн)	<a href="#">Обучающий</a> (13 млн)	<a href="#">Исторические</a> <sup>5</sup> (14 млн)
<a href="#">Социальные сети</a> (157 млн)	<a href="#">МультиПАРКи</a> <sup>2</sup> (458 тыс)	<a href="#">От 2 до 15</a> (4,4 млн)	<a href="#">Панхронический</a> (384 млн)

[все корпуса](#) <sup>49</sup>

Состав и структура

Статистика корпуса

Руководство пользователя

# Востребованность НКРЯ

☰ Google Академия

нкря



Статьи

Результатов: примерно 19 700 (0,08 сек.)



Мой профиль



Моя библиотека

☰ Google Академия

ruscorpora



Статьи

Результатов: примерно 16 200 (0,11 сек.)



Мой профиль



Моя библиотека

# Востребованность НКРЯ

*Для лингвистов НКРЯ сегодня  
является тем, чем для физиков  
адронный коллайдер.*

академик А. М. Молдован



# Основной сценарий использования

Лексико-грамматический поиск ? ↶

По умолчанию поиск ведется по предпочтительным разборам, слова могут совпадать.

Поиск учитывает:

- все разборы
- предпочтительные разборы

При поиске:

- совпадения слов исключаются
- слова могут совпадать

Слово 1 ⊗ + Слово 2 ⊗ +

Лемма ⊗ ?  
лекция  
добавить условие ▲

Грамм. признаки выбрать ? ⊗  
А

Расстояние ⊗  
от:  до:

Синтаксические отношения выбрать ? ⊗  
amod  
Направление связи  
 зависит  
 управляет  
от слова / словом  
Слово 1 ▲  
добавить условие ▲

Искать ▼ Сбросить

Запрос Вернуться к поиску • 1974 текста • 3604 примера предпочтительные разборы, сло...  
Добавить в сравнение

Конкорданс KWIC График Статистика Частотность 2-граммы 3-граммы 4-граммы 5-граммы Скачать ?

- 1. Думай, как хакер // «Computerworld», 2004** ⊗  
Слушателям будут прочитаны *подробные лекции* по методам проведения сетевых атак и защиты от них, базирующиеся на самой свежей информации в данной области. ⊗ ↶
- 2. Анна Фенько. Студент всегда прав // «Коммерсантъ-Власть», 2002** ⊗  
Предлагаем вместо зачета устроить *дополнительную лекцию*, а зачет поставить автоматом". ⊗ ↶
- 3. Андрей Геласимов. Фокс Малдер похож на свинью (2001)** ⊗  
Екатерине Михайловне нравились такие вещи. Однажды прочитала *целую лекцию*. Даже химия ослабила хватку. ⊗ ↶
- 4. Алексей Варламов. Купавна // «Новый Мир», 2000** ⊗  
Все в его байках выходило так непринужденно, завлекательно и ловко, будто Глеб сам в этой загадочной Сорбонне учился и на великолепном костюмированном балу танцевал; он дарил женщинам на Восьмое марта не мимозы, а сирень, рассказывал анекдоты про армянское радио, читал *экономические лекции* на коньячном заводе и

# Корпусная революция

*Корпусная революция в лингвистике очень сильно изменила саму лингвистику, её понимание... Корпус — это не просто инструмент, он меняет наши подходы к предмету изучения.*

академик В. А. Плунгян



# НКРЯ сегодня

## *Поиск по корпусам*

*Основной* (374 млн)

*Устный* (13 млн)

*Параллельные* <sup>28</sup> (179 млн)

*Поэтический* (13 млн)

*Газетные* <sup>2</sup> (850 млн)

*Акцентологический* (135 млн)

*Диалектный* (599 тыс)

*Русская классика*  $\beta$  (18 млн)

*Синтаксический* (1,6 млн)

*Мультимедийный* (5,8 млн)

*Обучающий* (13 млн)

*Исторические* <sup>5</sup> (14 млн)

*Социальные сети* (157 млн)

*МультиПАРКи* <sup>2</sup> (458 тыс)

*От 2 до 15* (4,4 млн)

*Панхронический* (384 млн)

# Специальные корпуса

## 20. А. А. Блок. Возмездие : «Жизнь — без начала и конца...» (1910-1921) !

Я4М Готòвься *лèкций* читать,  
Я4Ж Запóтанный в граждàнском прàве,  
Я4М С душòй, начàвшей ùставàть, —  
Я4Ж Он скрòмно прèдложил ей рúку,  
Я4М Связàл её с своей судьбой  
Я4М И в дàль увèз её с собой,  
Я4Ж Ужè питàя в сèрдце скúку, —  
Я4М Чтòбы женà с ним дò звездý  
Я4М Делýла книжныè трудý... 🗣️ ↔️



скачать

Penny (Kaley Cuoco) So/ you know/ isn't there maybe some way you and Sheldon could compromise on this whole presentation thing.

Leonard Hofstadter (Johnny Galecki) No. No. Scientists do not compromise. Our minds are trained to synthesise facts and come to inarguable conclusions. Not to mention/ Sheldon is batcrap crazy. 🗣️ ↔️



Пенни (Татьяна Шитова) Ну Ч

Лео́нард Хо́фстедтер (Андрей Фелдосов) Нет/ нет. Ученые не при компромиссам. Наши умы приучены сопоставлять факты и делать неоспоримые выводы. А Шелдон к тому же большой на всю голову. 🗣️ ↔️

## Гайдай, Морис Слободской, Яков Костюковский. Бриллиантовая рука, к/ф



Горбунов (Юрий Никулин) Вот я/ Варвара Сергевна/ был в Лондоне/ и там собаки гуляют везде. Собака/ друг *человека*.

Управдом (Нонна Мордюкова) А я не знаю/ как там в Лондоне. Я не была. Может там собака/ друг *человека*/ а у нас управдом/ друг *человека*.

### Жесты:

Имя говорящего (актера)	Пол говорящего (актера)	Активный орган	Название жеста	Значение жеста
Нонна Мордюкова	женский	голова	двигать головой вперед	подчеркивать ритм фразы
Игорь Ясулович	мужской	кисть	двинуть кисть к кому-л.	аргумент
Юрий Никулин	мужской	кисть	двинуть кисть к кому-л.	аргумент
Игорь Ясулович	мужской	кисть	держаться за сердце	нервозность
Нонна	женский	голова	качнуть головой	дистанцирование

### 1. О жизни (Приездово, Ржевский район, Тверская область, 1924) ⏪

Вòсшемдесеть *пять*. В авин тяплò накладàл. Пèсшни зжнàю, усй сштаринны успòмню. 🗣️ ↔️

### 2. О деревне, церкви, семье (Б. Ковали, Свечинский район, Кировская область, 1945) ⏪

Я работаю. Мне молока дают и яиц. До войны — то лучше было в городе жить. *Пять* месяцев жил. Не дают больно — то ловить. Не пушышай во свою сотню, Поликарповна. 🗣️ ↔️

# Корпус для переводчиков

## 21. Simon Sinek. Leaders Eat Last: Why Some Teams Pull Together and Others Don't (2014) | Саймон Сине́к. Лидеры едят последними. Как создать команду мечты (Е. И. Животикова, 2015) ⓘ

английский:

"The regulations that had kept finance boring had all but disappeared by the time Goldman's IPO was issued," wrote Harvard *Law* professor Lawrence Lessig in a column for CNN.com. ⓘ ↔

русский:

Как писал в своей колонке для CNN.com профессор *права* Гарвардского университета Лоуренс Лессиг: «Правила, регулировавшие финансовую часть, исчезли. ⓘ ↔

## 2. Simon Sinek. Leaders Eat Last: Why Some Teams Pull Together and Others Don't (2014) | Саймон Сине́к. Лидеры едят последними. Как создать команду мечты (Е. И. Животикова, 2015) ⓘ

английский:

Generation Y is said to have a sense of *entitlement*. ⓘ ↔

английский:

But I, as one observer, do not believe it is a sense of *entitlement*. ⓘ ↔

английский:

What we perceive as *entitlement* is, in fact, impatience. An impatience driven by two things: First is a gross misunderstanding that things like success, money or happiness, come instantly. ⓘ ↔

русский:

У поколения Y было чувство собственного *права*. ⓘ ↔

русский:

Но я, как сторонний наблюдатель, не верю, что это чувство собственного *права*. ⓘ ↔

русский:

То, что мы принимаем за чувство собственного *права*, по сути, является нетерпением, вызванным двумя вещами. Во-первых, это грубое заблуждение, будто такие вещи, как успех, деньги или счастье, можно получить сразу же. ⓘ ↔

английский:

And there's a structure that sits on the left and the *right* side of your brain, called the hippocampus. ⓘ ↔

английский:

I believe it is now time for us to reclaim our *right* to a full night of sleep, and without embarrassment or that unfortunate stigma of laziness. ⓘ ↔

английский:

MW: So you're *right*, we can't catch up on sleep. ⓘ ↔

русский:

На слайде вы можете наблюдать гипоталамус, расположенный в *правом* и левом полушариях мозга. ⓘ ↔

русский:

Я считаю, что пришло время вернуть наше *право* на полноценный сон, не испытывая стыда и не боясь прослыть лентяями в глазах общества. ⓘ ↔

русский:

МУ: Да, ты *права*, мы не можем «наверстать» сон. ⓘ ↔

) | Мэтт Уолкер. Сон — это ваша суперсила

# Корпус для учителей

## Упражнения на основе Корпуса ? <

Значение слова (один). Морфология. Частеречная омонимия (один)

Классы: 6 Темы: лексическая семантика, морфология



Исторические изменения значений слов (эк и эка)

Классы: 9 Темы: лексикология, лексическая семантика



Исторические изменения значений слов («покойник»)

Классы: 9 Темы: лексикология, лексическая семантика



Синонимия («пламя» и «огонь»)

Классы: 6 Темы: лексикология, лексическая семантика



Переносное значение. Метафора («рыцарь»)

Классы: 6 Темы: лексикология, лексическая семантика



Многозначность («белый»)

Классы: 6 Темы: лексикология, лексическая семантика



Значение. Многозначность. Эпидигматика («предложение»)

Классы: 10-11 Темы: лексикология, лексическая семантика



### Темы

Лексикология

Лексическая семантика

Морфемика

Словообразование

Морфология

показать ещё ↓

### Классы

5

5-6

5-7

6

6-7

показать ещё ↓

### Сложность

Олимпиадное

Применить

Очистить

# Корпус для NLP-инженеров

## Скачиваемые корпуса:

- Морфологический стандарт
- Синтаксический датасет
- Диахронический датасет
- Многоязычный датасет
- Датасет n-грамм Основного корпуса

## Нейромоделли

- Токенизатор
- Векторные модели
- Морфемные модели
- Модели метаразметки

# НКРЯ и МЛ

# Почему корпусу нужны ML-based инструменты?

1. НКРЯ слишком велик (как по числу документов, так и по числу слов) для разметки руками.
2. С использованием новой разметки возникают новые инструменты поиска.

НейроКРЯ



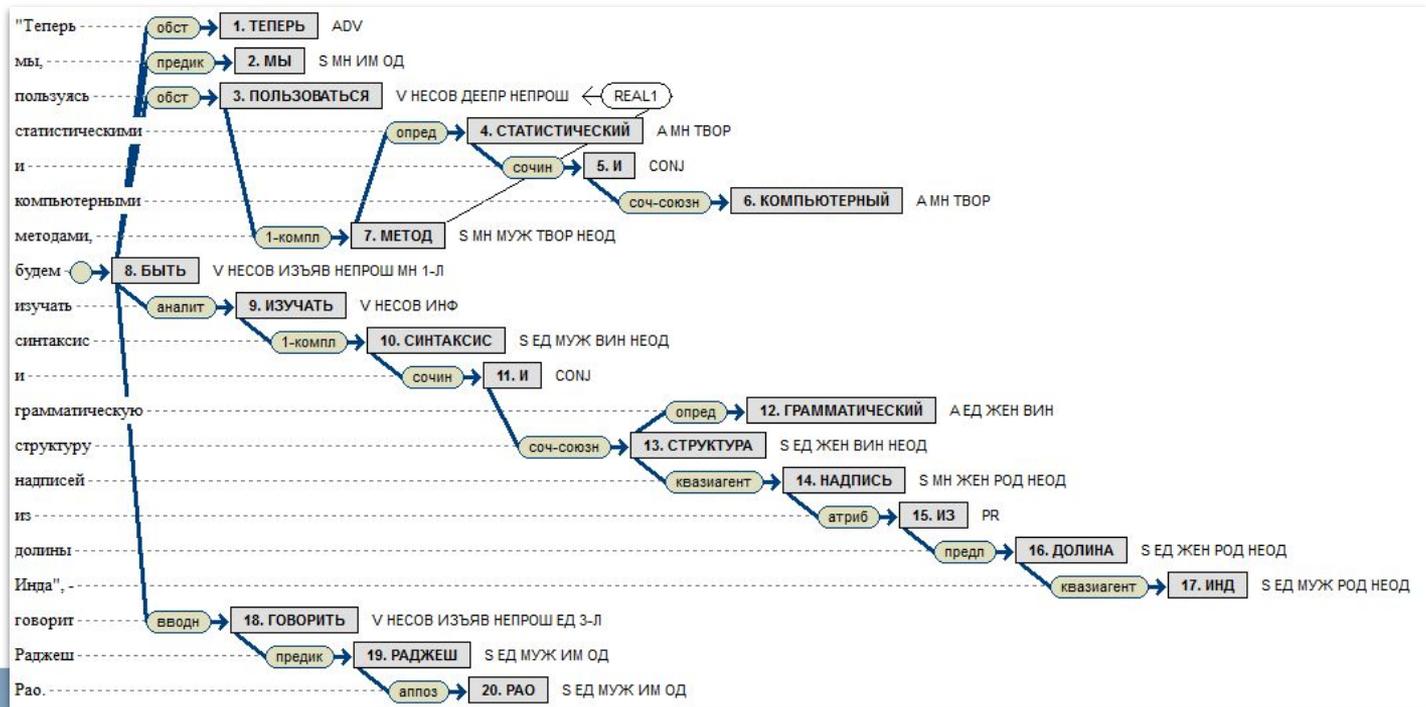
# РуБик: разметка лемм и морфологии

LSTM-энкодер с тремя декодерами: для морфологии, лемм и синтаксиса.  
Набор правил постпроцессинга.

Качество:

- Части речи: **99.03%**
- Лемматизация: **98.79%** (в следующей версии — **99.07%**)
- Полная морфология: **97.27%**
- Синтаксическое дерево без учёта типа связей: **95.08%**
- Синтаксическое дерево с учётом типа связей: **93.64%**

# РуБик: разметка синтаксиса



# Синтаксические скетчи

<b>биология</b> Существительное					
<i>Определения</i>		<i>Сказуемые</i>		<i>Глаголы с прямым дополнением</i>	
1. молекулярный	11,73	1. развиваться	4,07	1. преподавать	7,15
2. мичуринский	10,4	2. учить	3,83	2. изучать	5,12
3. эволюционный	8,35	3. требовать	2,06	3. изучить	4,88
4. синтетический	7,97	4. давать	1,25	4. сдавать	4,72
5. физико-химический	7,86	5. работать	1,18	5. учить	3,31
6. экспериментальный	7,76	6. остаться	0,62	6. сдать	2,97
7. клеточный	7,55	7. оказаться	0,57	7. касаться	1,57
8. космический	6,95			8. создать	1,39
9. лысенковский	6,95			9. поставить	1,07
10. радиационный	6,82			10. знать	0,39

# Разные парадигмы разметки морфем



насекомое



насекомое

# Морфемная модель — ансамбль CNN

Морфемный разбор β ?

★ Оценить

сгенерировано НейроКРЯ

эстетика

Однокоренные слова β ?

★ Оценить

сгенерировано НейроКРЯ

эстетика  
эстетический  
эстет  
эстетически  
эстетизм  
эстетик  
эстетский  
неэстетичный  
эстетизация  
эстетизировать  
неэстетично

# НКРЯ и эволюция русского языка

# Тысяча лет русского языка

Распределение результатов поиска по датам (частота на миллион словоформ) с 1000 по 2021 ?

Статистика рассчитана с учетом совпадающих слов

Детализация по годам

Период с: 1021



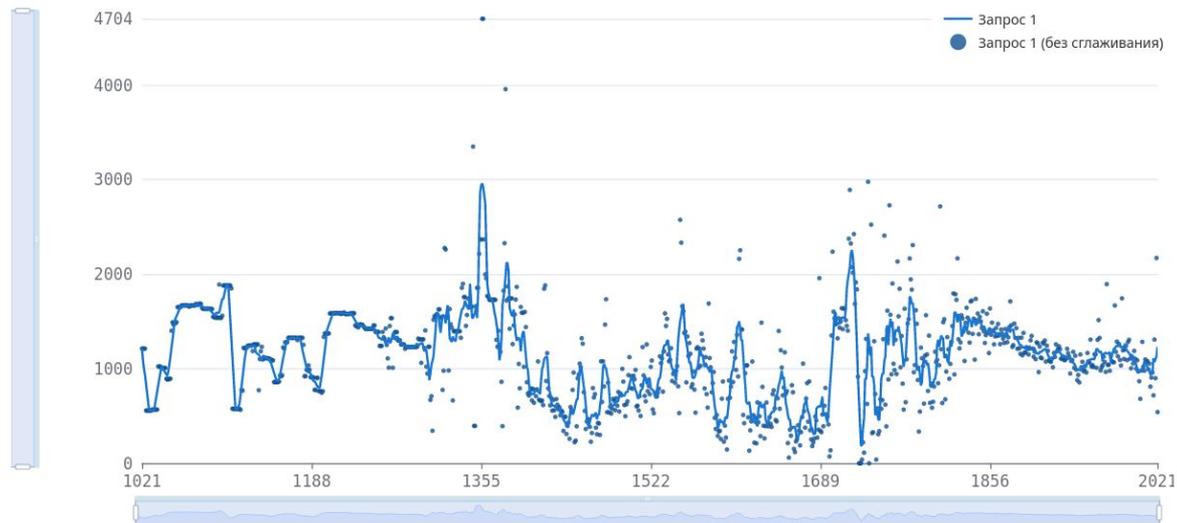
по: 2021



со сглаживанием 3

Построить

Данных за пределами 1100-2021 может быть не достаточно для статистически значимых выводов



Искать в [Google Books Ngram Viewer](#)

# Корпус берестяных грамот



531

+ ѿ аны покло ко климате брате господине пощелоуи о моемо  
 орууде коснатини  
 оу а нине извета емоу людемн како еси возложило поруку на  
 мою сестру и на дочерь еи и  
 а зовело еси сестру мою кривою и дочере блядею а нинеца  
 феоде прехудоу оуслышаво то слово  
 и быгноло сестру мою и хотело потати а нинеца господине  
 брате согдаво со воеславом мовоу емоу  
 тако еси возложило то слово таково дроведи аже ти возомови  
 коснатини дала роуку  
 (а) за зате ты же браце господине мовоу емоу тако  
 оже боудоу люди на мою сестру оже боудоу люди при комо  
 боудоу дала роуку за зате то те а во вине  
 ты покю брате испытыво которе слово звало на ма и порукою а  
 боудоу люди на томо тобе не сестра  
 а мовужени не жена ты же ма и потени не зера на федора и  
 дала ма доци коуны людемн сызвето  
 мо а закладо просила и позовало мене во погосто и а зю прехала  
 оже оно поехало проше а река тако  
 а зю соло 'а: дворано по гривене сыра

531 об.

оже вождоу людемн на мовоу сестру оже боудоу люди на  
 ты пакко берлати спзати во кооторе слово звало нинеца  
 а мовужени не жена ты же ма и потени не зера на федора и  
 дала ма доци коуны людемн сызвето  
 мо а закладо просила и позовало мене во погосто и а зю прехала  
 оже оно поехало проше а река тако  
 а зю соло 'а: дворано по гривене сыра

0 5 10 CM

## 1. Берестяная грамота 531 (1200-1220)

### Оригинал:

+ ѿ аны покло ко климате брате господине пощелоуи о моемо  
 орууде коснатини  
 оу а нине извета емоу людемн како еси возложило поруку на  
 мою сестру и на дочерь еи и  
 а зовело еси сестру мою кривою и дочере блядею а нинеца  
 феоде прехудоу оуслышаво то слово  
 и быгноло сестру мою и хотело потати а нинеца господине  
 брате согдаво со воеславом мовоу емоу  
 тако еси возложило то слово таково дроведи аже ти возомови  
 коснатини дала роуку  
 (а) за зате ты же браце господине мовоу емоу тако  
 оже боудоу люди на мою сестру оже боудоу люди при комо  
 боудоу дала роуку за зате то те а во вине  
 ты покю брате испытыво которе слово звало на ма и порукою а  
 боудоу люди на томо тобе не сестра  
 а мовужени не жена ты же ма и потени не зера на федора и  
 дала ма доци коуны людемн сызвето  
 мо а закладо просила и позовало мене во погосто и а зю прехала  
 оже оно поехало проше а река тако  
 а зю соло 'а: дворано по гривене сыра

русский

От Анны поклон Климате. Господин брат, вступишь за меня  
 перед Коснатином в моем деле. Объяви ему при свидетелях:  
 «После того как ты возложил поручительскую  
 ответственность ((букв.): поручительство) на мою сестру и на  
 ее дочь ((т. е.) заявил, что они поручились [и] назвал сестру  
 мою кривою, а дочь блядью, теперь Фед (Федор), приехавши  
 и услышав об этом обвинении, выгнал сестру мою и хотел  
 убить». Так что, господин брат, согласовавши с Воеславом,  
 скажи ему (Коснатиному): «[Раз] ты предъявил это обвинение,  
 так докажи». Если же скажет Коснатиин: «Она поручилась за  
 зятя», — то ты, господин братец, скажи ему так: «Если будут  
 свидетели против моей сестры, если будут свидетели, при  
 ком она ((букв.:) я) поручилась за зятя, то вина на ней ((букв.:)  
 на мне)». Когда же ты, брат, проверишь, какое обвинение и  
 [какое] поручительство он (Коснатиин) на меня взвел, то, если  
 найдутся свидетели, подтверждающие это, — я тебе не  
 сестра, а мужу не жена. Ты же меня и убей, не глядя на  
 Федора ((т. е.) не принимая его во внимание). А дала моя  
 дочь деньги при людях, с публичным объявлением и  
 требовала заклада. А он (Коснатиин) вызвал меня в погост, и я  
 приехала, потому что он уехал со словами: «Я шлю четырех  
 дворян за гривнами серебра ((т. е.) чтобы они взяли  
 положенный штраф)».

# Словарь неправильностей

**В. ДОЛОПЧЕВЪ.**

—

**ОПЫТЪ СЛОВАРЯ**

**НЕПРАВИЛЬНОСТЕЙ**

**ВЪ РУССКОЙ РАЗГОВОРНОЙ РѢЧИ**

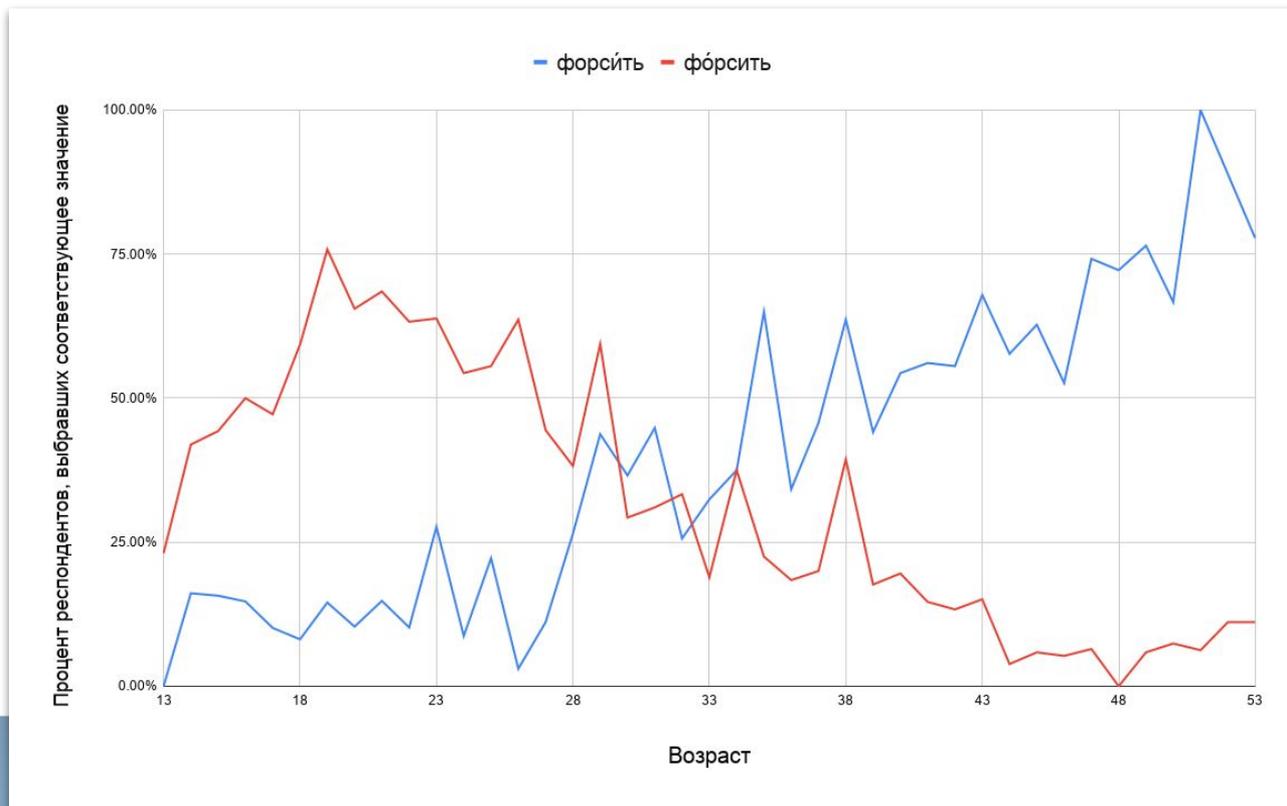
-- —

# Словарь неправильностей

**Обыденный** (н. зн. однодневный, въ теченіе одного дня сдѣланный, однѣ сутки дѣющійся: *Обыденная церковь* въ Москвѣ и Вологдѣ, по преданію, построенная въ однѣ сутки; *обыденный путь* — что можно пройти или проѣхать въ сутки; *обыденный мотылекъ* — живущій однѣ сутки) — обиходный, вседневный. *Въ обыденной жизни. Въ простомъ обыденномъ платьѣ. Обыденный случай.*

*Прим.* Противъ несообразнаго употребленія слова **обыденный** особенно возстаютъ знатоки языка В. Даль и Я. Гротъ

# Любите ли вы форсить?



# Коварные слова

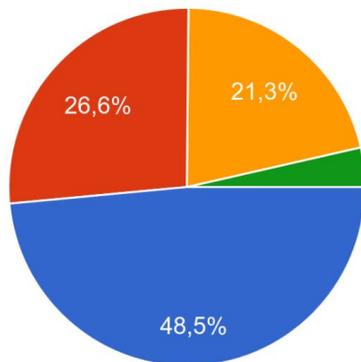
Слова, которые кажутся носителю языка знакомыми, но, отвечая на уточняющий вопрос о значении, человек ошибается.

Примеры найденных нами коварных слов: зябь, конгениальный, органичный, нелицеприятный, изморось.

# Коварные слова

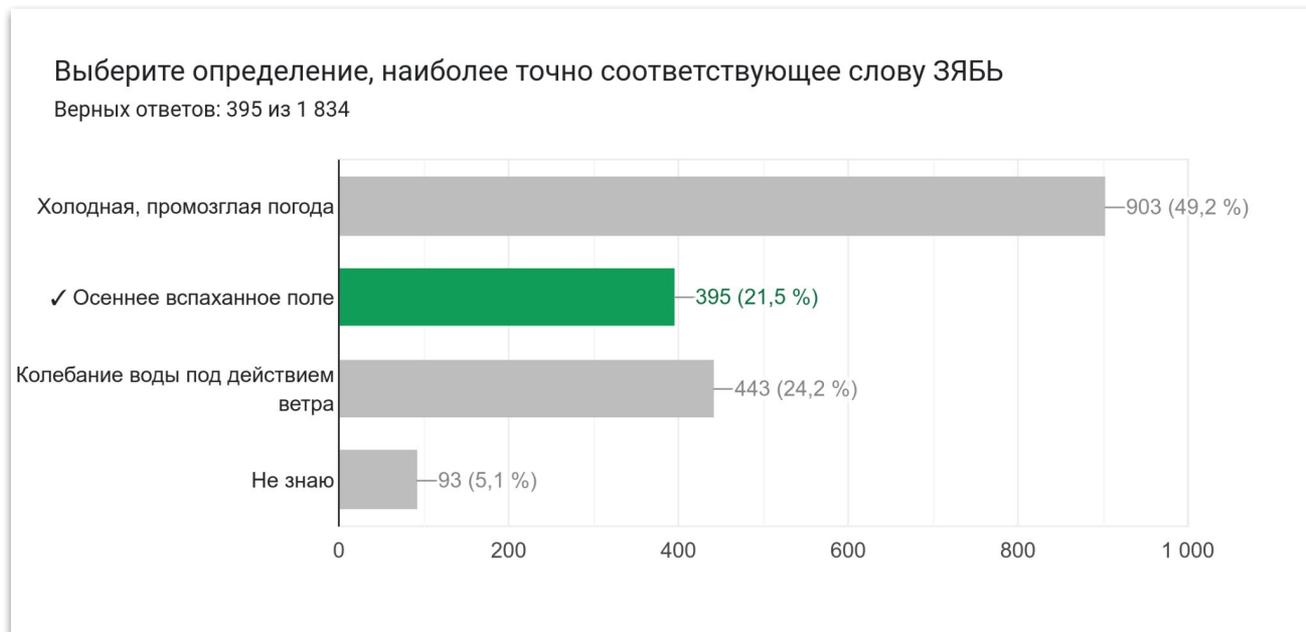
Знакомо ли Вам слово «зябь»?

169 ответов



- Хорошо знакомо, мог(ла) бы объяснить его своими словами
- Неплохо знакомо, помню контекст, в котором оно употребляется
- Видел(а) когда-то, но не помню точно
- Вижу это слово впервые

# Коварные слова



# Автоматический поиск семантических сдвигов — RuShiftEval'21

Соревнование по автоматическому поиску семантических сдвигов.

111 существительных, 3 диахронических датасета: досоветский, советский и постсоветский.

Лучшее решение — на XML-R при помощи адаптации алгоритма, решающего задачу WSD — разделения разных значений одного слова.

# Поиск новых значений слов

Толковые словари достаточно быстро устаревают, а составление такого словаря — огромный человеческий труд. Можно ли помочь экспертам? Да, можно!

Генерация толкований — пока что недостаточное качество (но постоянно улучшается, LLM, вероятно, скоро смогут).

Можно найти такие контексты, в которых слово встречается в несловарном значении.

GlossBERT — бинарный классификатор для пары (контекст, толкование).  
Получили  $F1 = 0.96$ .

# Диахронические word2vec-модели: «???»

до 1800: столбик; зазубрина; подкладка; дощечка; фас

1800-1825: дровни; высушить; ковать; мигом; черенок

1825-1850: кабриолет; дрожки; тележка; колымага; палочка

1850-1875: передок; тележка; перекладина; санки; палочка

1875-1900: повозка; фура; фургон; дрога; телега

1900-1920: тележка; фура; двуколка; дрога; арба

1920-1940: панель; передок; ролик; жердочка; козлы

1940-1960: ролик; станина; циркуль; рейка; катушка

1960-1980: циркуль; столбик; ролик; логарифмический; угольник

1980-2000: циркуль; рейка; столбик; пенал; дощечка

2000-2010: выкройка; циркуль; лекало; стежок; плоттер

2010-2020: тренажер; модуль; комплект; клавиатура; трек

# Диахронические word2vec-модели: «линейка»

до 1800: столбик; зазубрина; подкладка; дощечка; фас

1800-1825: дровни; высушить; ковать; мигом; черенок

1825-1850: кабриолет; дрожки; тележка; колымага; палочка

1850-1875: передок; тележка; перекладина; санки; палочка

1875-1900: повозка; фура; фургон; дрога; телега

1900-1920: тележка; фура; двуколка; дрога; арба

1920-1940: панель; передок; ролик; жердочка; козлы

1940-1960: ролик; станина; циркуль; рейка; катушка

1960-1980: циркуль; столбик; ролик; логарифмический; угольник

1980-2000: циркуль; рейка; столбик; пенал; дощечка

2000-2010: выкройка; циркуль; лекало; стежок; плоттер

2010-2020: тренажер; модуль; комплект; клавиатура; трек

# Диахронические word2vec-модели: «???»

до 1800: площадь; сторона; торжище; остров; кладбище

1800-1825: базар; сено; лавка; хата; мельница

1825-1850: базар; биржа; ярмарка; толкучий; чердак

1850-1875: базар; биржа; ярмарка; толкучий; торг

1875-1900: базар; товар; торг; торговля; фабрика

1900-1920: базар; биржа; товар; ярмарка; торговец

1920-1940: базар; толкучка; товар; биржа; торговля

1940-1960: базар; толкучка; ярмарка; магазин; товар

1960-1980: базар; толкучка; торг; ярмарка; товар

1980-2000: биржа; торговля; товар; базар; ярмарка

2000-2010: биржа; спрос; экспорт; цена; импорт

2010-2020: спрос; биржа; экспорт; продажа; инвестор

# Диахронические word2vec-модели: «рынок»

до 1800: площадь; сторона; торжище; остров; кладбище

1800-1825: базар; сено; лавка; хата; мельница

1825-1850: базар; биржа; ярмарка; толкучий; чердак

1850-1875: базар; биржа; ярмарка; толкучий; торг

1875-1900: базар; товар; торг; торговля; фабрика

1900-1920: базар; биржа; товар; ярмарка; торговец

1920-1940: базар; толкучка; товар; биржа; торговля

1940-1960: базар; толкучка; ярмарка; магазин; товар

1960-1980: базар; толкучка; торг; ярмарка; товар

1980-2000: биржа; торговля; товар; базар; ярмарка

2000-2010: биржа; спрос; экспорт; цена; импорт

2010-2020: спрос; биржа; экспорт; продажа; инвестор

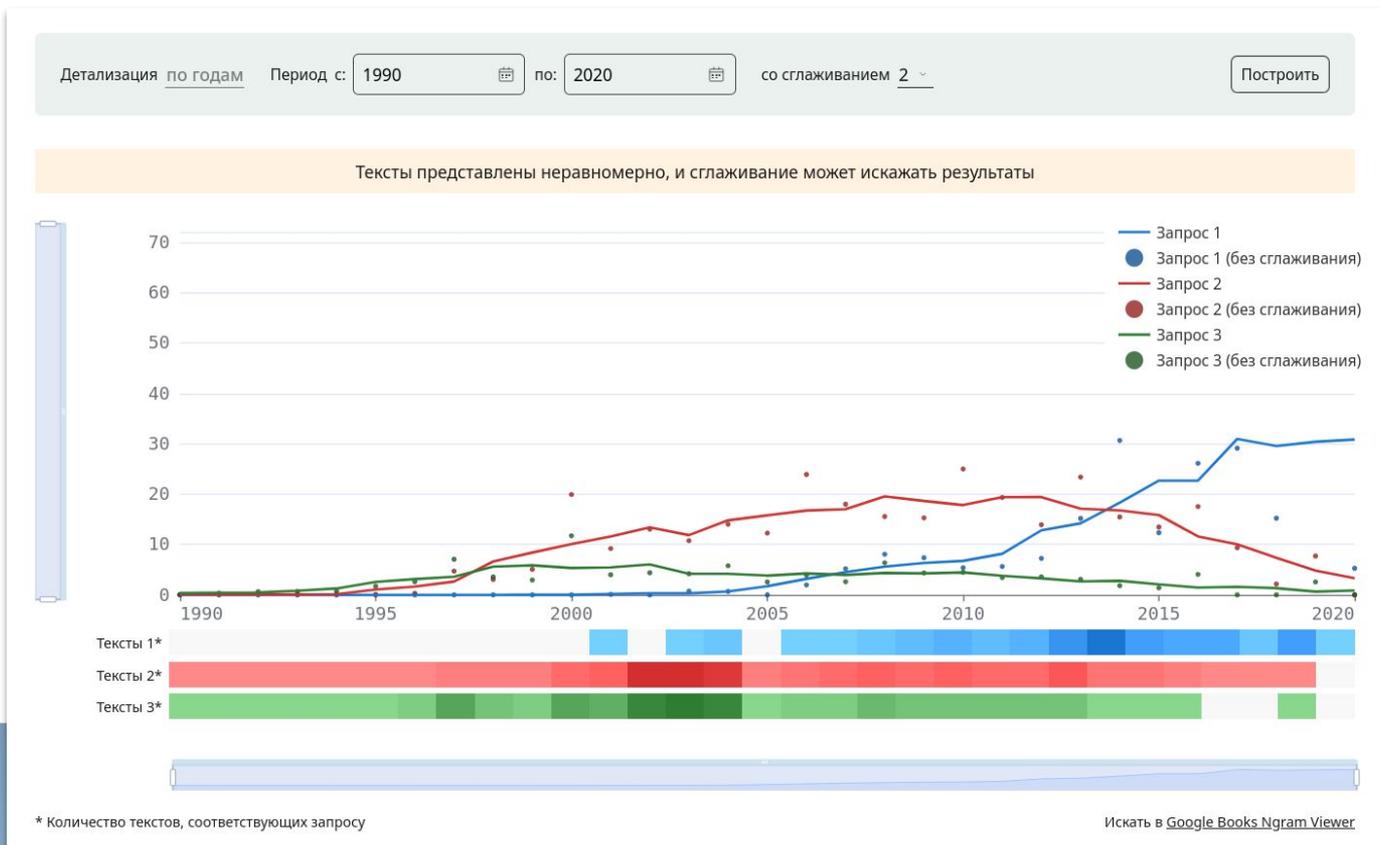
# Диахронические word2vec-модели: «рубль»

До 1800: червонный; копейка; левк; ефимок; пуд  
1800-1825: червонец; пиастр; червонный; золотый; талер  
1825-1850: талер; франк; копейка; червонец; червонный  
1850-1875: франк; целковый; доллар; копейка; талер  
1875-1900: франк; иена; талер; доллар; целковый  
1900-1920: иена; франк; доллар; талер; крона  
1920-1940: франк; доллар; червонец; динар; иена  
1940-1960: франк; доллар; фунт; копейка; марка  
1960-1980: доллар; франк; копейка; лира; цент  
1980-2000: доллар; франк; шекель; копейка; афгани  
2000-2010: доллар; гривна; евро; бакс; йена  
2010-2020: доллар; евро; бакс; гривна; фунт

# А что кроме семантики?

оувѣдѣти    послоушати    разгнѣватисѧ  
повѣдати    съказати  
обѣщатисѧ    **СЛЫШАТИ**    оуслышати  
възвѣстити    оубоитисѧ  
видѣти

# Употребимость синонимов



# Употребимость синонимов

Запросы [Графики](#)



	<a href="#">Запрос 1</a>	<a href="#">Запрос 2</a>	<a href="#">Запрос 3</a>
	<b>Запрос</b>		
Поиск учитывает:	предпочтительные разборы	предпочтительные разборы	предпочтительные разборы
При поиске:	соседние слова могут совпадать	соседние слова могут совпадать	соседние слова могут совпадать
	<b>Слово 1</b>		
Лемма	смартфон	мобильный	сотовый
	<b>Слово 2</b>		
Лемма	-	телефон	телефон
Расстояние	-	на расстоянии от 1 до 1 от Слова 1	на расстоянии от 1 до 1 от Слова 1

# Портрет слова

# Портрет слова

портрет 🗨️ ? Часть речи: Любая (часть речи) Показать портрет

Существительное

Скетчи ?

портрет Существительное

Определения	Сказуемые	Глаголы с прямым
1. скульптурный 9,06	1. висеть 10,19	1. нарисовать
2. фотографический 8,98	2. красоваться 7,32	2. рисовать
3. акварельный 8,64	3. храниться 6,91	3. повесить
4. фамильный 8,45	4. украшать 6,49	4. писать
5. словесный 8,43	5. изображать 6,35	5. написать
6. групповой 8,4	6. сохраниться 5,9	6. заказать
7. поясной 8,3	7. удаться 5,73	7. набросать
8. миниатюрный 8,05	8. понравиться 5,71	8. поместить
9. карандашный 7,7	9. передавать 5,21	9. подарить
10. живописный 7,65	10. помещаться 5,18	10. вешать

Показать все коллокации

Формы слова ?

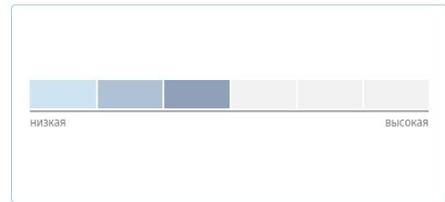
Падеж	единственное	множественное
именительный	портрет	портреты
родительный	портрета	портретов
дательный	портрету	портретам
винительный	портрет	портреты
творительный	портретом	портретами
предложный	портрете	портретах

О слове ?

портрет

**Лемма** портрет (с.м. словарь)  
**Грамматика** существительное, неодушевленное, мужской  
**Семантика** предметные имена, произведение изобразительного искусства

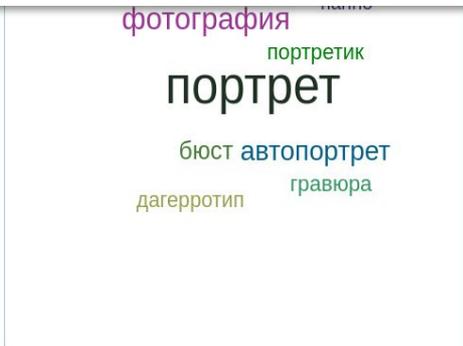
Частотность слова ?



Морфемный разбор ? Оценить



Однокоренные слова ? Оценить



# Портрет слова

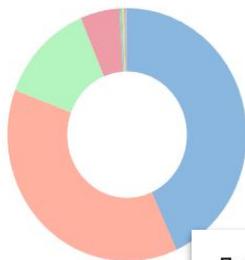
## Статистика текстов ?

Метаатрибут

Сфера функционирования

Показывать

слова



- публицистика (43.19%)
- художественная (37.56%)
- бытовая (12.93%)
- учебно-научная (5.33%)
- электронная коммуникация (0.35%)
- церковно-богословская (0.24%)
- реклама (0.18%)

## Примеры ?

Большое широкое лицо, довольно длинно обстриженные волосы, небольшой рост, неприятное выражение глаз, все черты несимпатичные — вот *портрет* человека, о котором так долго говорила вся Европа во время войны 1877-1878 годов.

Наверно, два боевых ордена, которые успеваешь рассмотреть читатель на груди комиссара, это немало, но, если бы внимательный читатель не успел их узреть, все равно было бы впечатление о Пантелееве как о человеке сильном и цельном, — в *портрете*, который написал писатель, это присутствует.

Особенно интересны его *портреты* исторических и научных деятелей.

Лучшие образцы этих работ на выставке — кинжалы, пояса, рамки для фотографий, гозыри для черкески, портсигары, кинжалы и дощечки с черным *портретом* Шота Руставели в обрамлении из узоров, исполненных техникой «грехилируи» работы Джикия, — отправлены на Парижскую выставку.

Никто до Белинского не умел так ясно видеть в произведениях Шекспира галерею живых *портретов*, а также «игру взаимных отношений и интересов всех лиц».

Показать все примеры

## Распределение результатов поиска по датам (частота на миллион словоформ) ?

Статистика рассчитана с учетом совпадающих слов

Детализация по годам

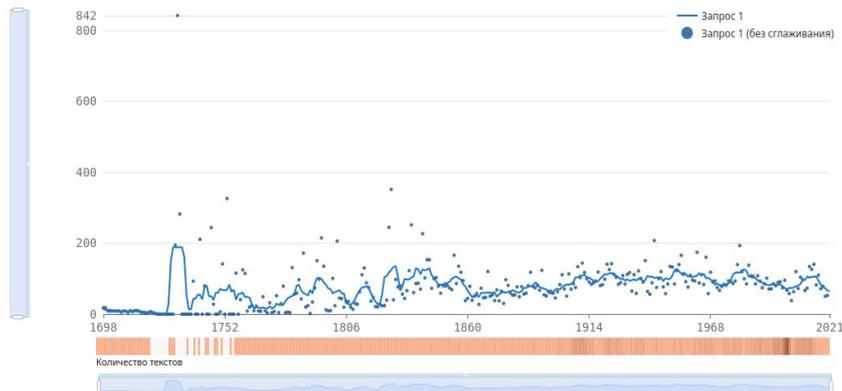
Период с: 1698

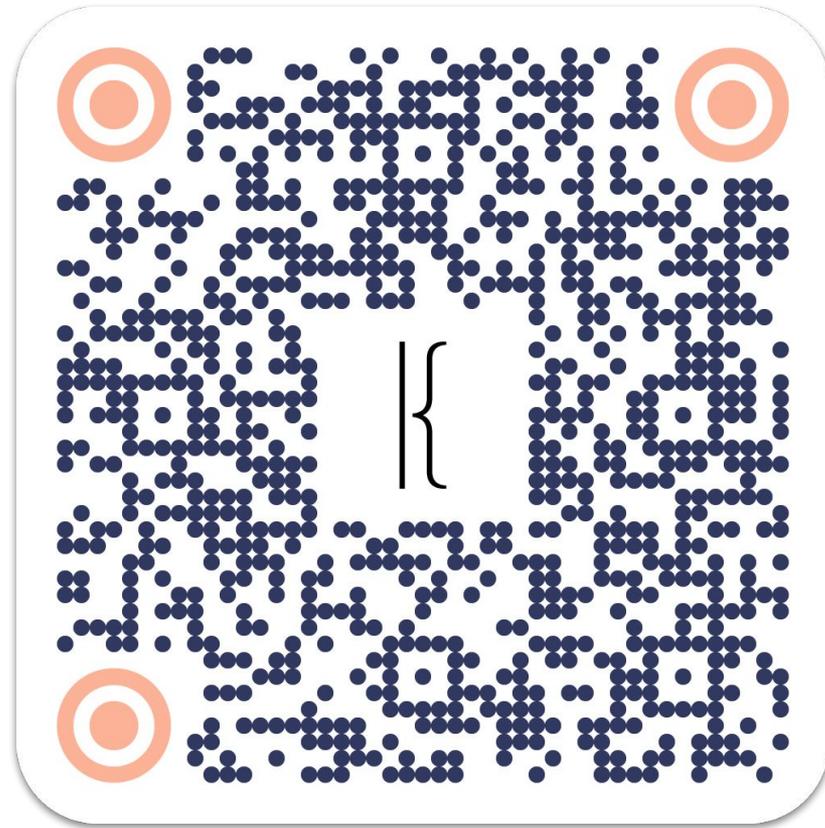
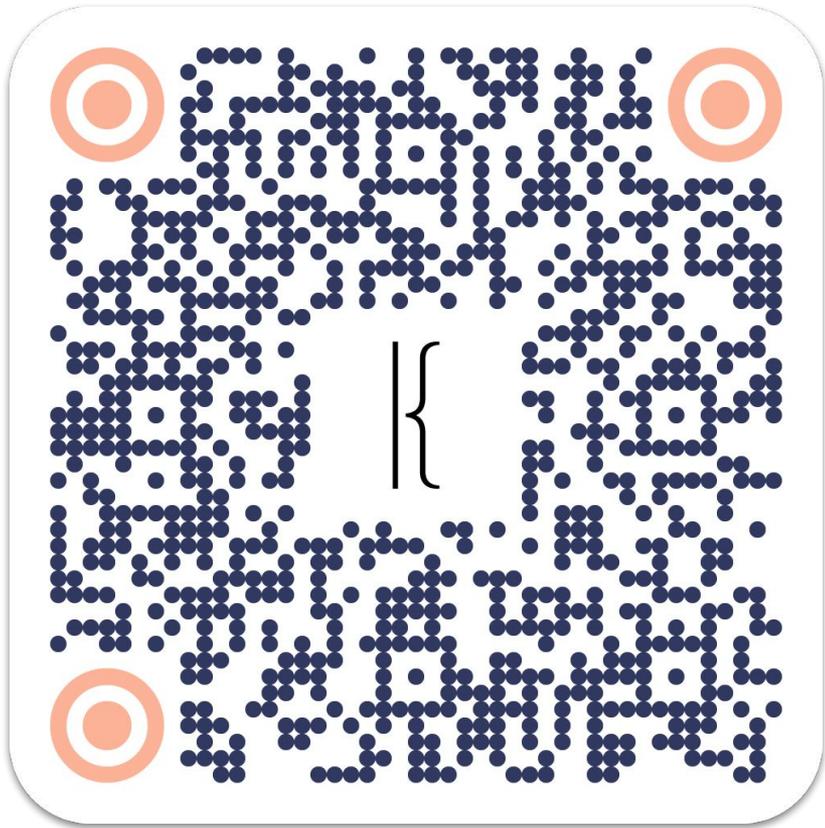
по: 2021

со сглаживанием 3

Построить

Тексты представлены неравномерно, и сглаживание может искажать результаты





# Спасибо за внимание!

---

Дмитрий Морозов

Лаборатория прикладных цифровых технологий ММЦ НГУ  
Национальный корпус русского языка

[morozowdm@gmail.com](mailto:morozowdm@gmail.com) | [t.me/morozowdm](https://t.me/morozowdm)



