Как корпусная лингвистика помогает изучать эволюцию языка

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

Russian NATIONAL CORPUS

mathematics & mechanics

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 корпусы?

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 корпусы?

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

Зачем нужны корпусы?

Фактически, любой корпус является виртуальным носителем языка.

Задачи, решаемые при помощи корпусов:

- изучение языка лингвистами
- проверка лингвистических гипотез при редактуре текста
- преподавание языка как родного и как иностранного
- обучение переводчиков
- обучение языковых моделей, 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

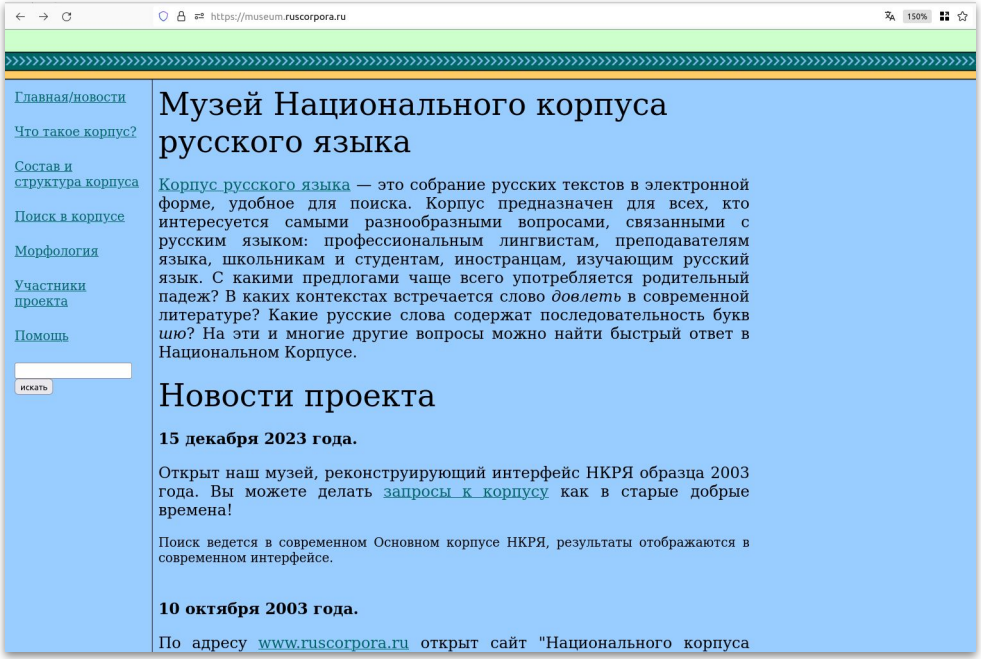
Сейчас, в 2005 г. мы должны признать, что данное научное направление (информатизация русистики) оказалось нежизнеспособным в современных организационно-финансовых условиях и постановка задачи создания Машинного фонда русского языка на ближайшую перспективу должна быть еще более сужена до двух-трех частных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

В. М. Андрющенко

Национальный корпус русского языка

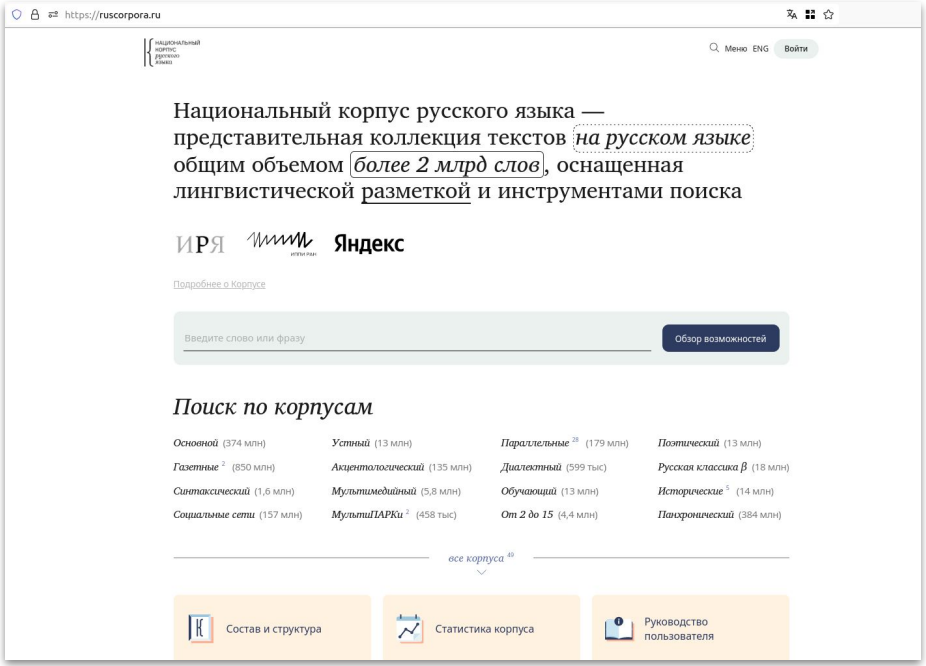
Корпус с начала 2000-х собирался и размечался командой лингвистов, а программировали корпусный движок специалисты из Яндекса — И. В. Сегалович и В. А. Титов.

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!



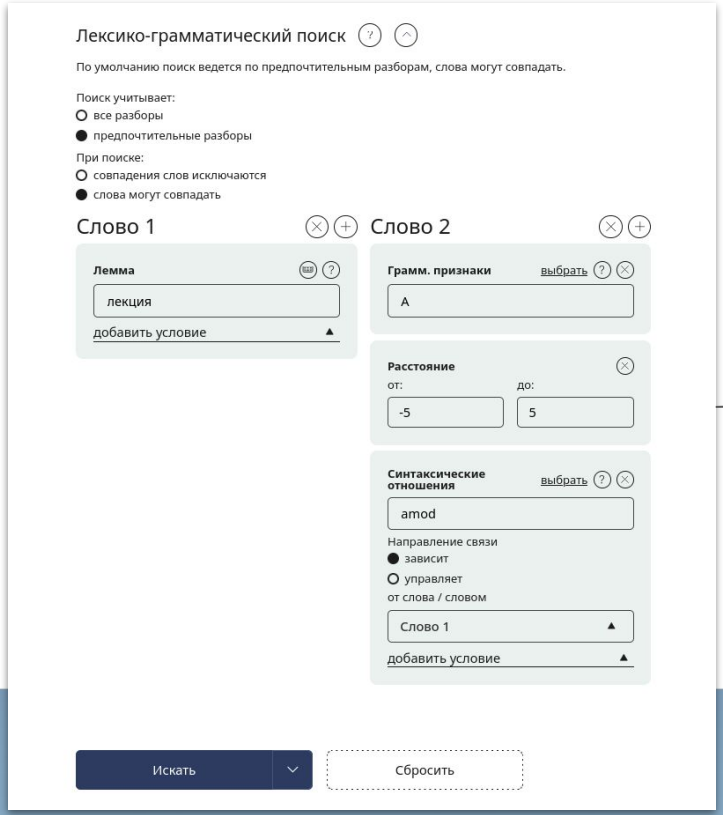
Востребованность НКРЯ

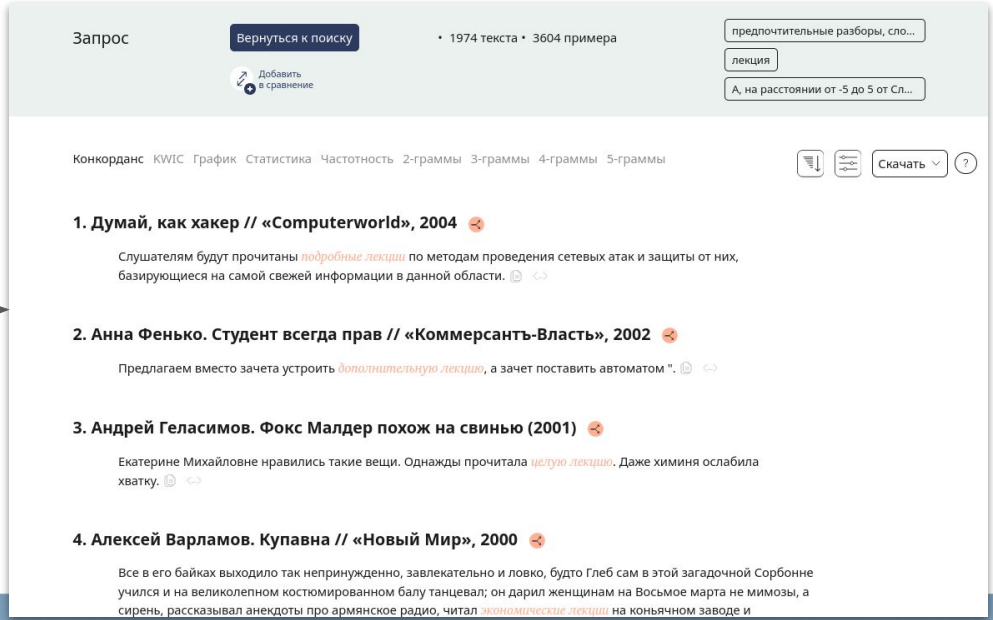
Востребованность НКРЯ

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

академик А. М. Молдован

Основной сценарий использования





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* [28] (179 млн)	*Поэтический* (13 млн)
Газетные [2] (850 млн)	*Акцентологический* (135 млн)	*Диалектный* (599 тыс)	*Русская классика β* (18 млн)
Синтаксический (1,6 млн)	*Мультимедийный* (5,8 млн)	*Обучающий* (13 млн)	*Исторические* [5] (14 млн)
Социальные сети (157 млн)	*МультиПАРКи* [2] (458 тыс)	*От 2 до 15* (4,4 млн)	*Панхронический* (384 млн)

Специальные корпусы

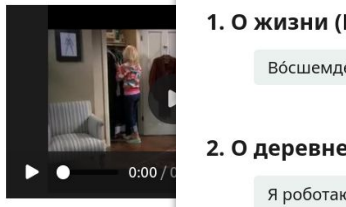
20. А. А. Блок. Возмездие : «Жизнь — без начала и конца...» (1910-1921)

Я4м Гото̀вясь *лѐкциѝ* чита̀ть,
Я4ж Запу̀танны̀й в граждà нском пра̀ве,
Я4м С душо̀ѝ, нача̀вшей у̀става̀ть, —
Я4ж Он скро̀мно прѐдложѝл ей ру̀ку,
Я4м Связа̀л еѐ с своѐй судьбо̀ѝ
Я4м И в да̀ль увѐз еѐ с собо̀ѝ,
Я4ж Ужѐ пита̀я в сѐрдце ску̀ку, —
Я4м Чтобы̀ жена̀ с ним до̀ звезды̀
Я4м Делѝла кнѝжныѐ труды̀...

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
придёте к компромисс
Леонард Хофстедтер (Андрей Федосов) Нет/ нет. Ученые не при
компромиссам. Наши умы приучены сопоставлять факты и делать неоспоримые выводы. А Шелдон к тому же больной на всю голову.

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

Горбунков (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

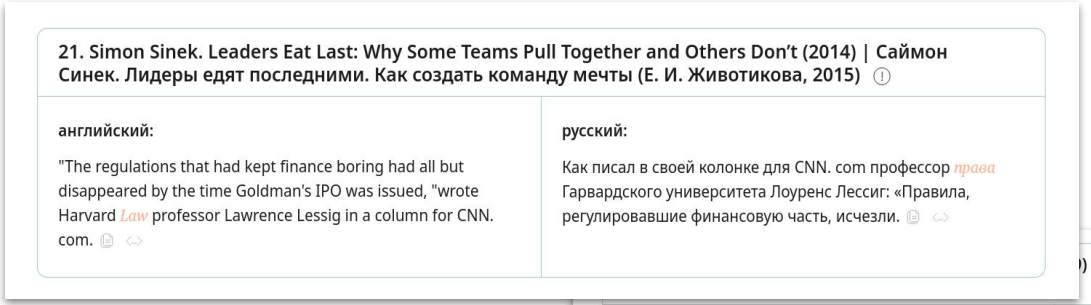
1. О жизни (Приездово, Ржевский район, Тверская область, 1924)

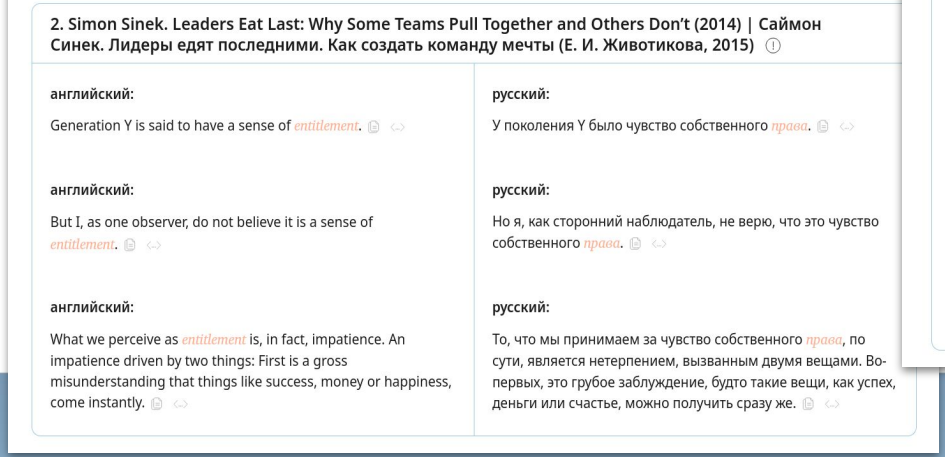
Во́сшемдесьъть *пя́ть*. В ави́н тяпло́ наклада́л. Пе́сшни зжна́ю, уси́ сштари́нны успо́мню.

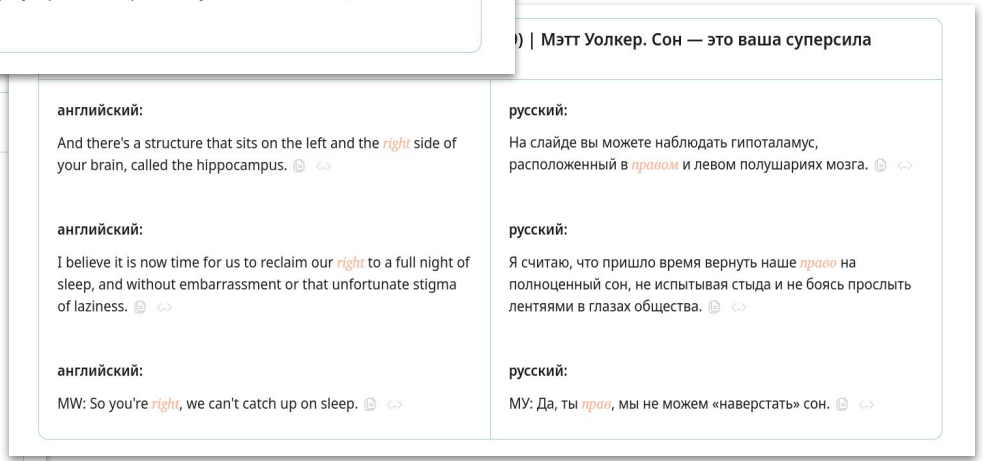
2. О деревне, церкви, семье (Б. Ковали, Свечинский район, Кировская область, 1945)

Я роботаю. Мне молока дают и яич. До войны — то лутьче было в городе жить. *Пять* месечов жил. Не дают больно — то ловить. Не пушьшяй во свою сотьню, Поликарповна.

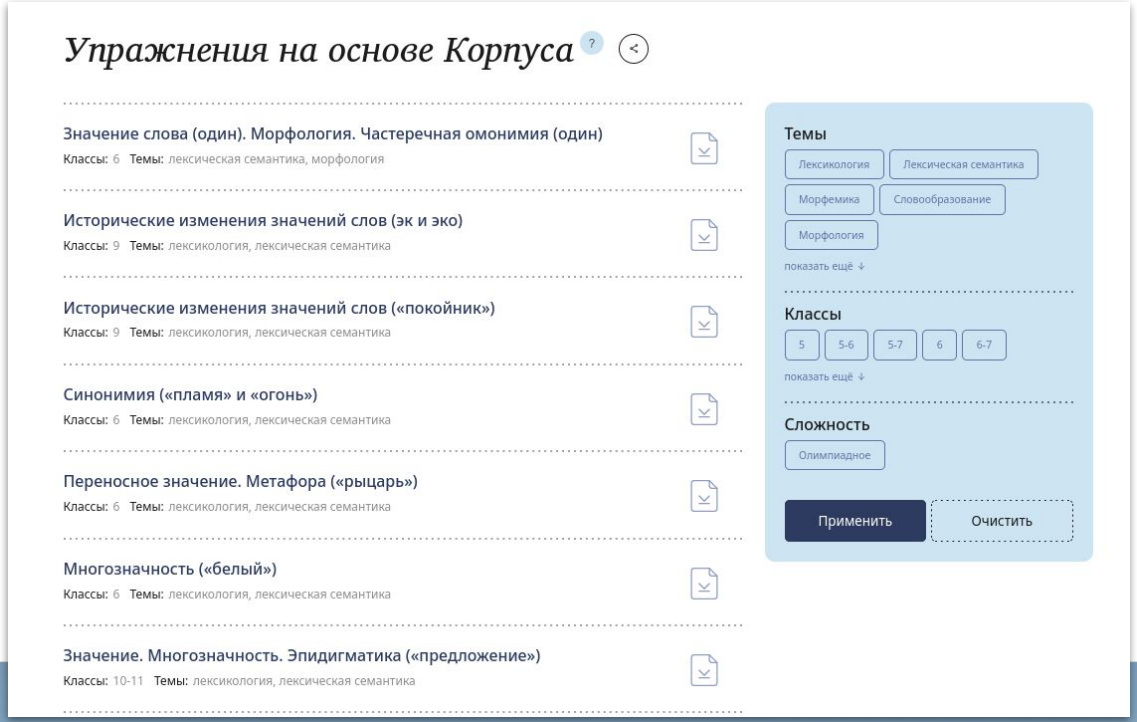
Корпус для переводчиков







Корпус для учителей



Корпус для NLP-инженеров

Скачиваемые корпуса:

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

Нейромодели

- Токенизатор
- Векторные модели
- Морфемные модели
- Модели метаразметки

НКРЯ и ML

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

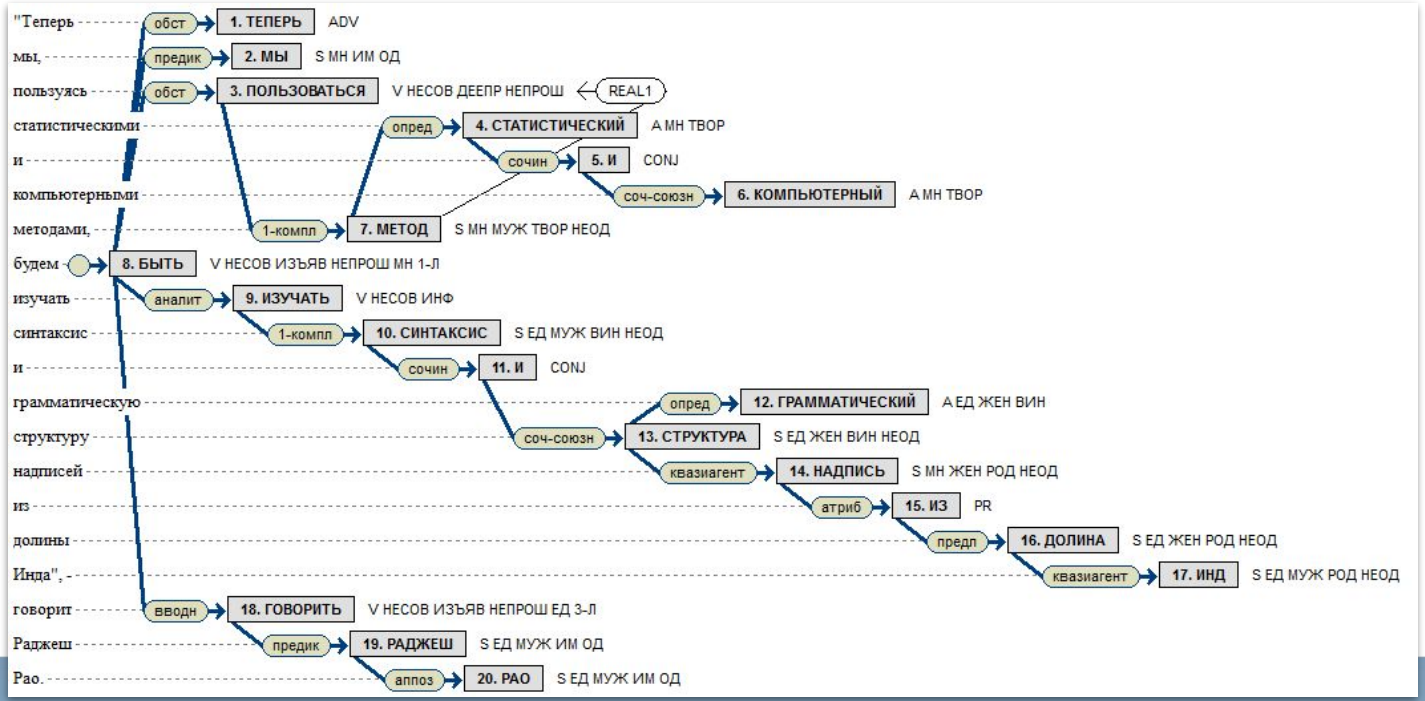
РуБик: разметка лемм и морфологии

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



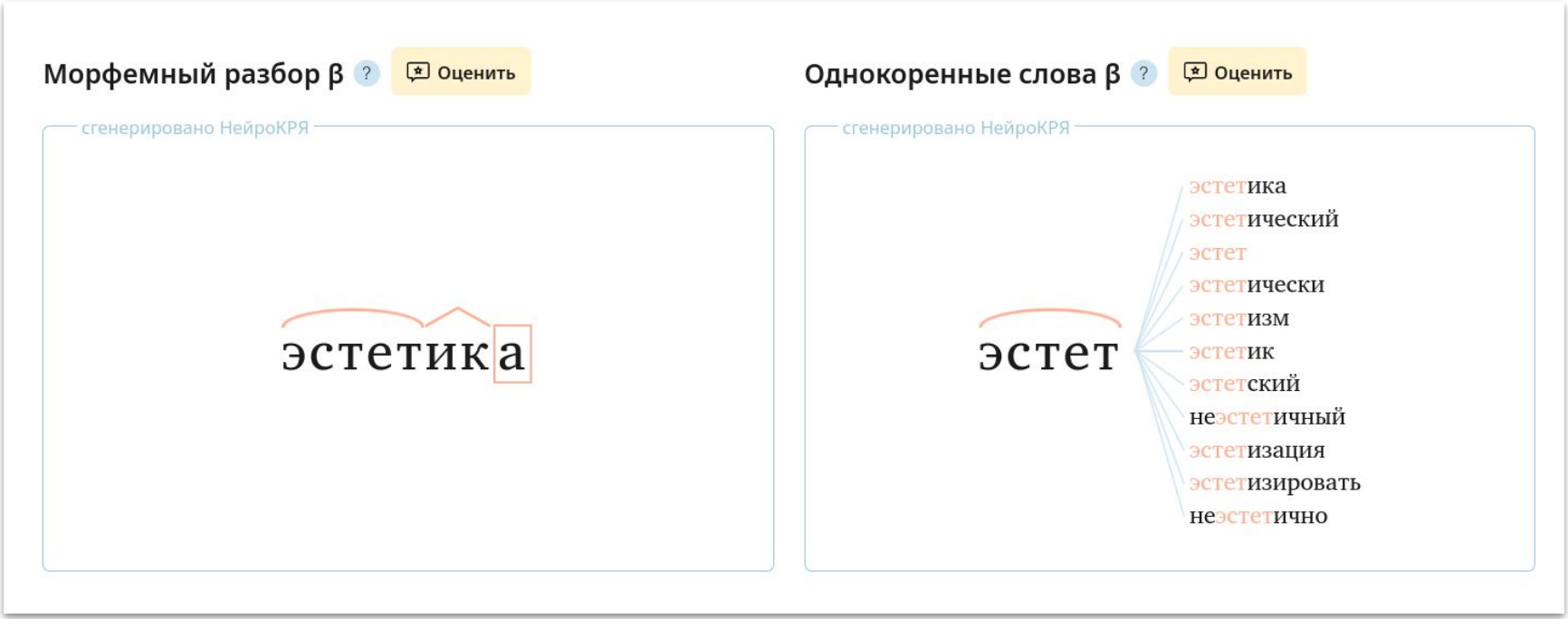
Синтаксические скетчи

биология Существительное

Определения		*Сказуемые*		*Глаголы с прямым дополнением*	
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

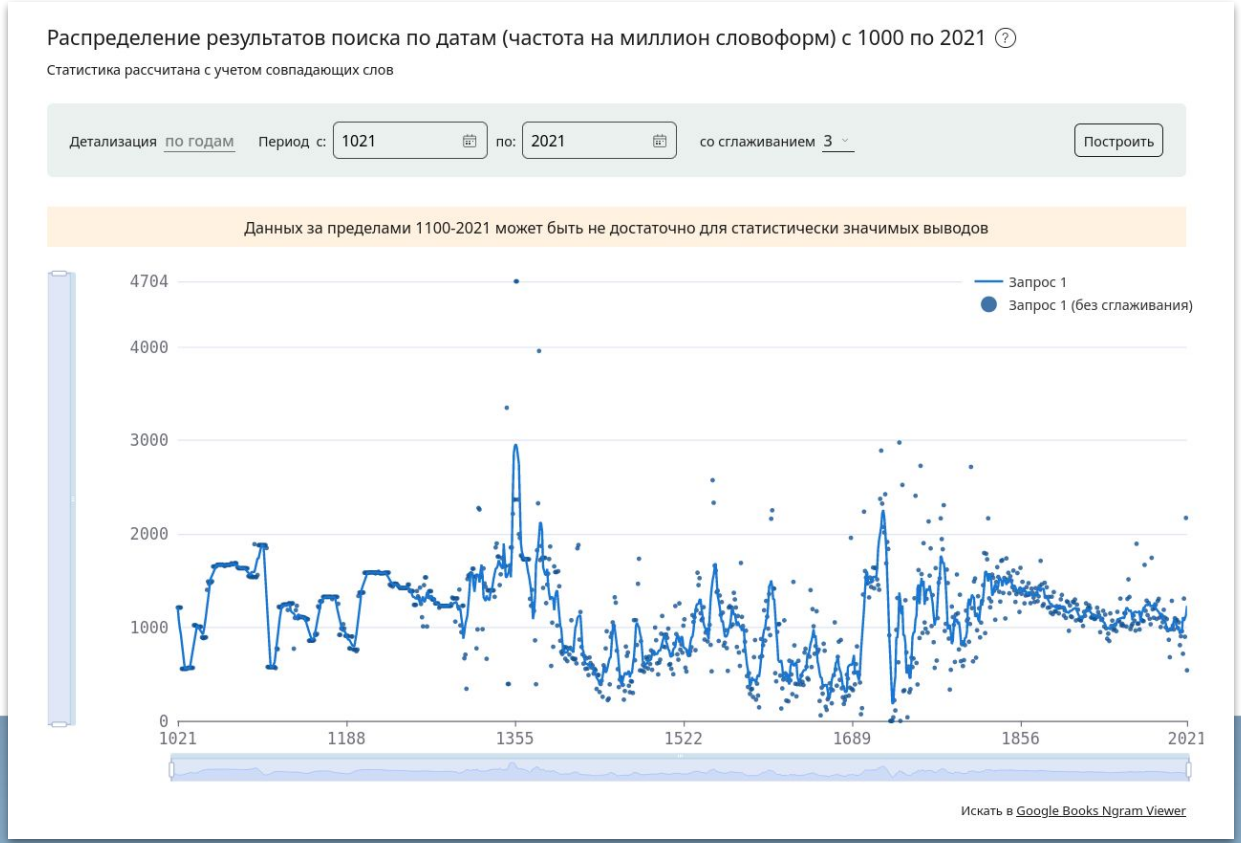
Разные парадигмы разметки морфем

Морфемная модель — ансамбль CNN

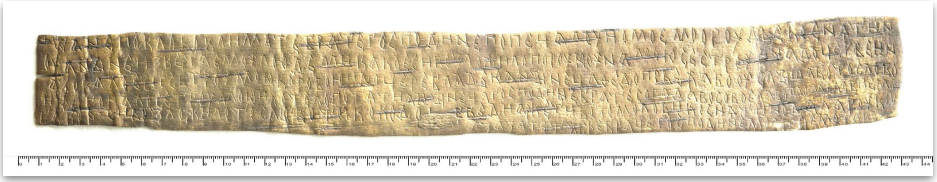


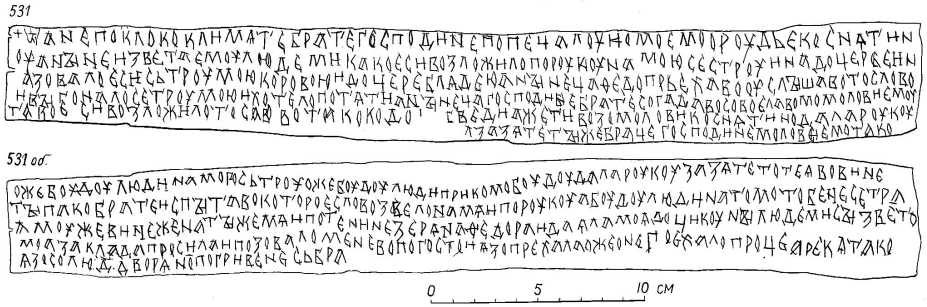
НКРЯ и эволюция русского языка

Тысяча лет русского языка



Корпус берестяных грамот





1. Берестяная грамота 531 (1200-1220)

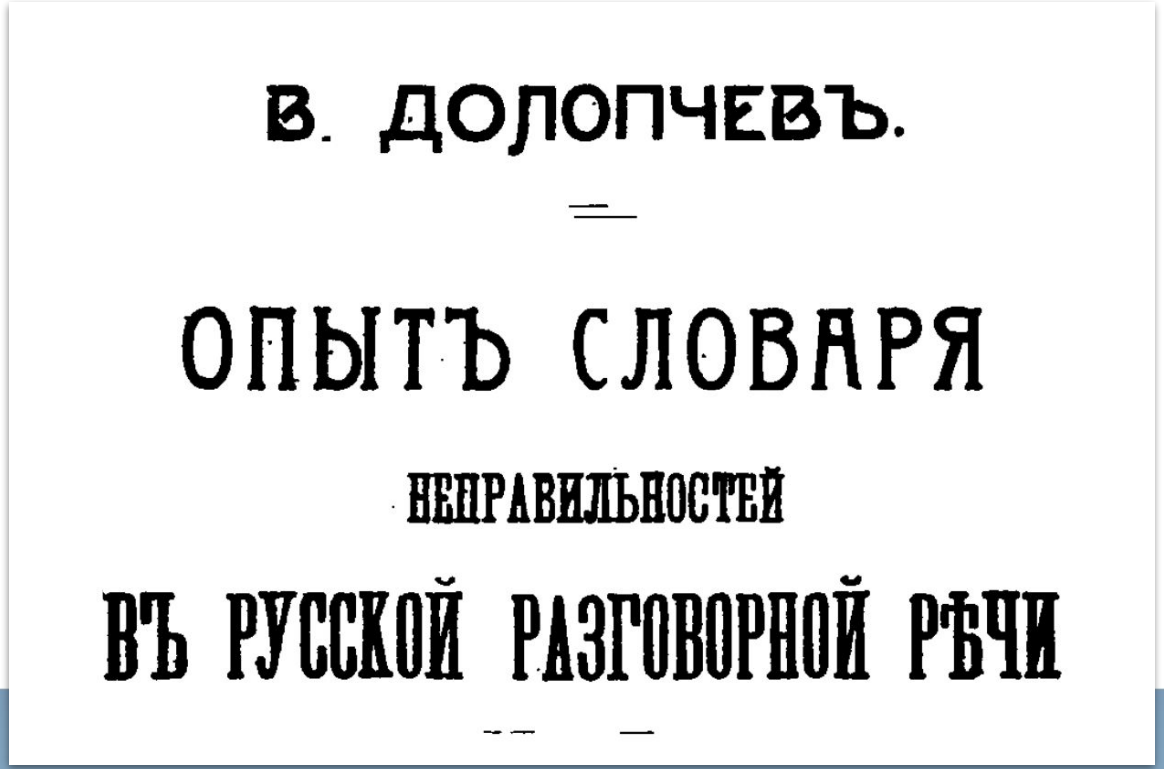
Оригинал:

+ ѿ дне покло ко климате брате господине попецалоуи о моемо
ѧроудье коснатин҃
оу ѧ мѧне изветѣ емоу людеми како еси возложило поручкоу на
мою сестроу и на дощерь еи и҃
ѧзовало еси сѣтроу мою коровою и дощере бладею ѧ мѧнеца
ѳедо прьехаво оуслъшиво то слово
и въгонало сетроу мою и хотело потати ѧ мѧнеца господине
брате согадаво со воеславомо молови емоу
тако еси возложило то слово тококо доведи ѧже ти возмолови
коснатино дола роукоу
{л} за зате тъı же браце господине молови емо тако
оже боудоу люди на мою сѣтроу оже боудоу люди при комо
боудоу дола роукоу за зате то те ѧ во вине
тъı пако брате испътаво которое слово звело на ма и поруκоу ѧ
боудоу люди на томо тобе не сетра
ѧ моужеви не жена тъı же ма и потени не зера на ѳедора и
дола моа доци коумъı людеми с ъзвето҃
мо ѧ заклада просило и позволо мене во погосто и ѧзо прехала
оже оно поехало проце ѧ река тако
ѧзо солю ⁘ дворано по гривене сьбра

русский

От Анны поклон Климяте. Господин брат, вступись за меня перед Коснятином 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 курвою, 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 (Коснятину): «[Раз] ты предъявил это обвинение, так докажи». Если же скажет Коснятин: «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 (Коснятин) 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 давала моя дочь деньги при людях, с публичным объявлением и требовала заклада. А он (Коснятин) 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

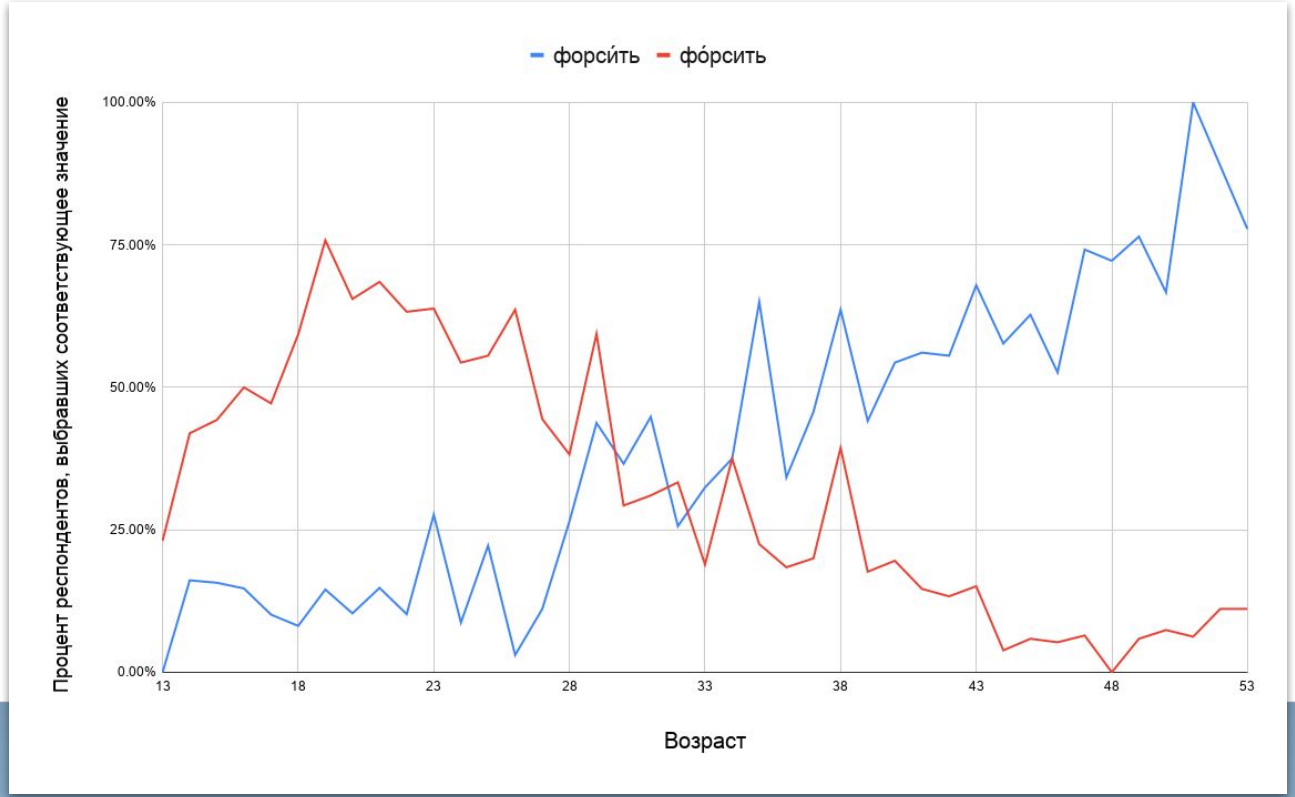
ВЪ РУССКОЙ РАЗГОВОРНОЙ РѢЧИ

Словарь неправильностей

Обыденный (н. зн. однодневный, въ теченіе одного дня сдѣланный, однѣ сутки длящійся: *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, повседневный. *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?

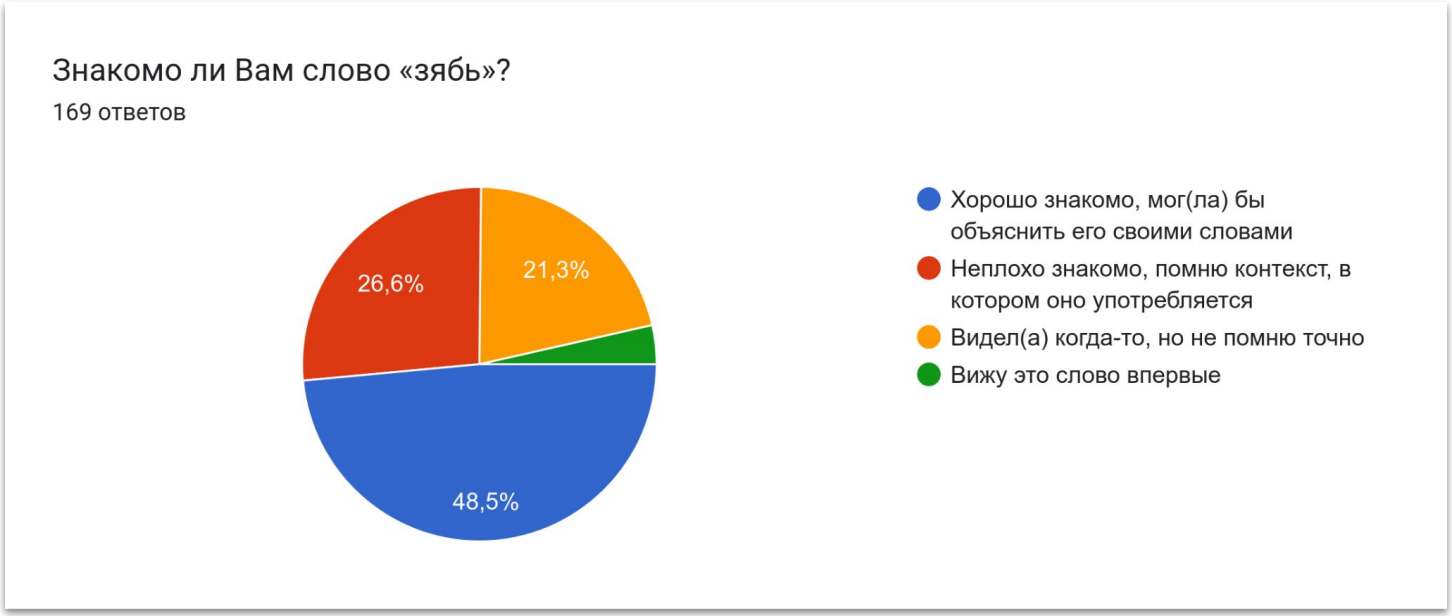


Коварные слова

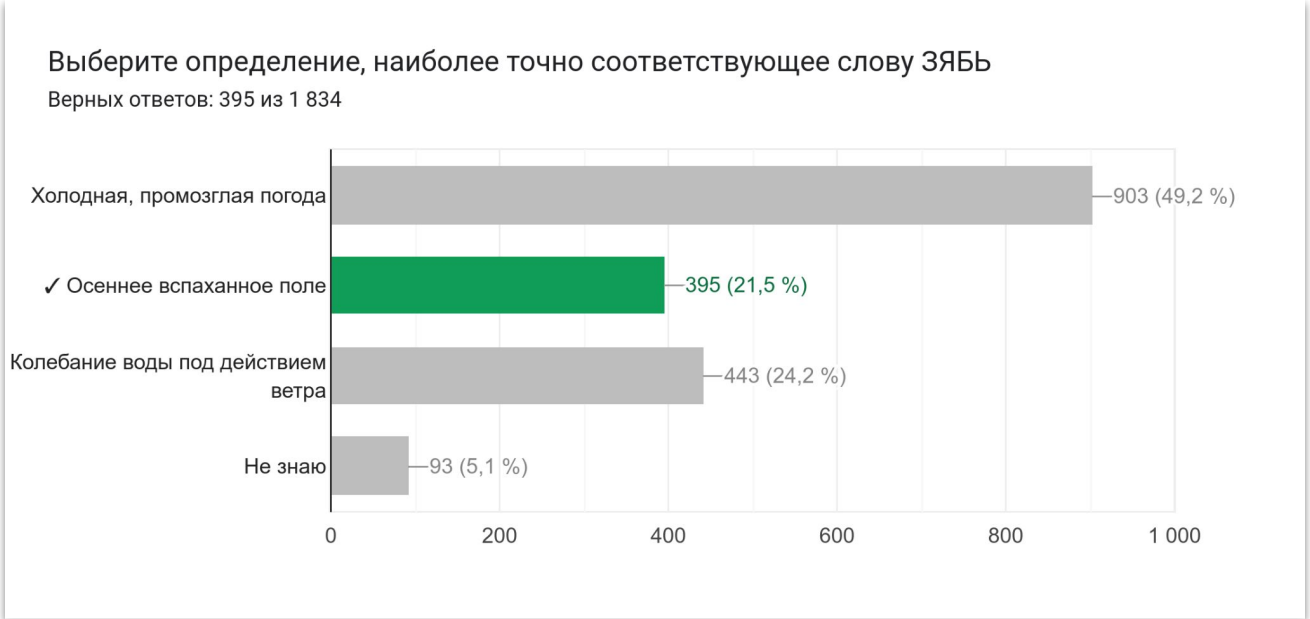
Слова, которые кажутся носителю языка знакомыми, но, отвечая на уточняющий вопрос о значении, человек ошибается.

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

Коварные слова



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

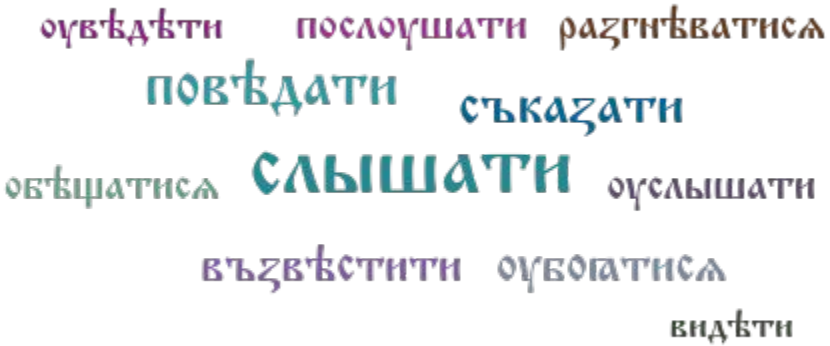
Диахронические word2vec-модели: «рынок»

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

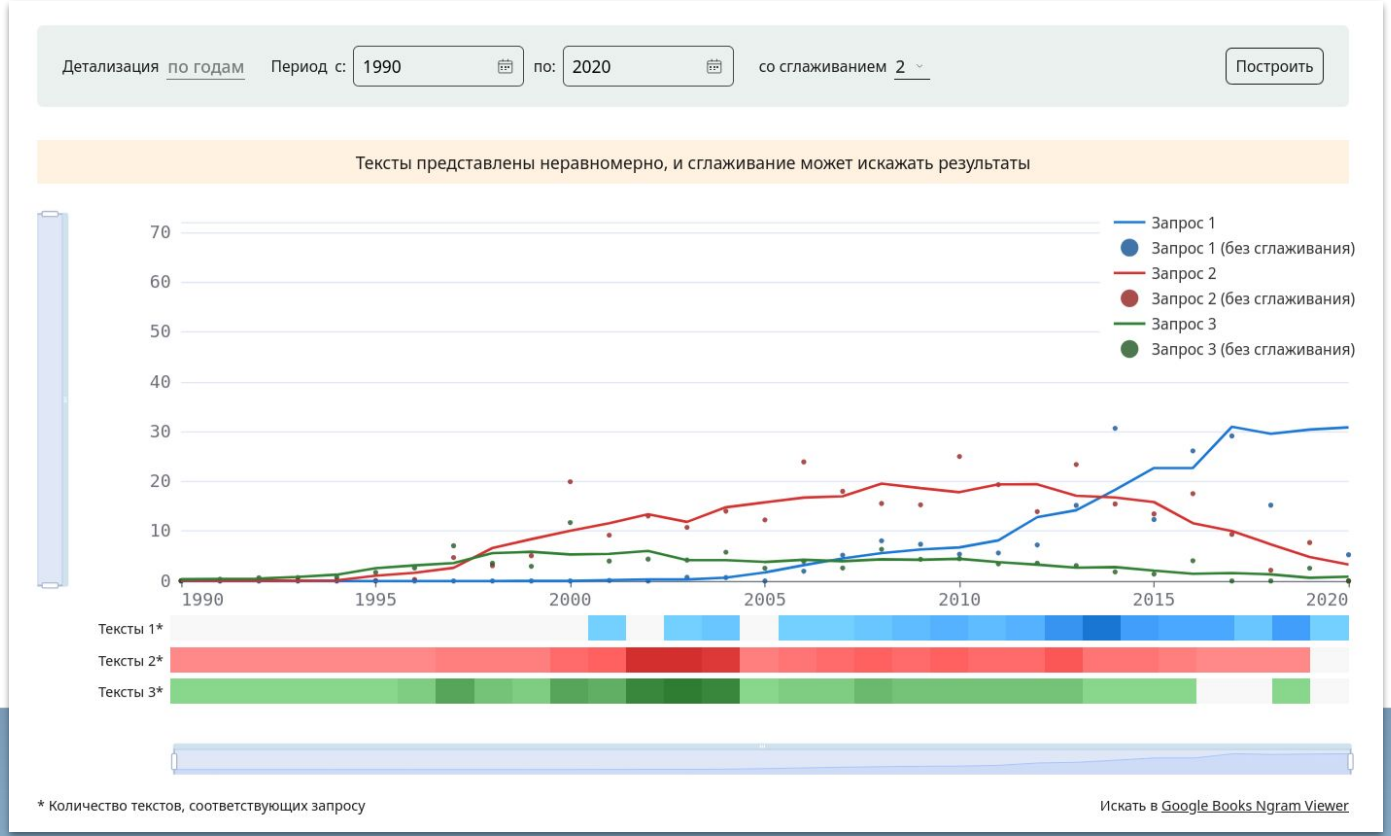
Диахронические word2vec-модели: «рубль»

До 1800: червонный; копейка; левк; ефимок; пуд
1800-1825: червонец; пиастр; червонный; злотый; 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?



Употребимость синонимов



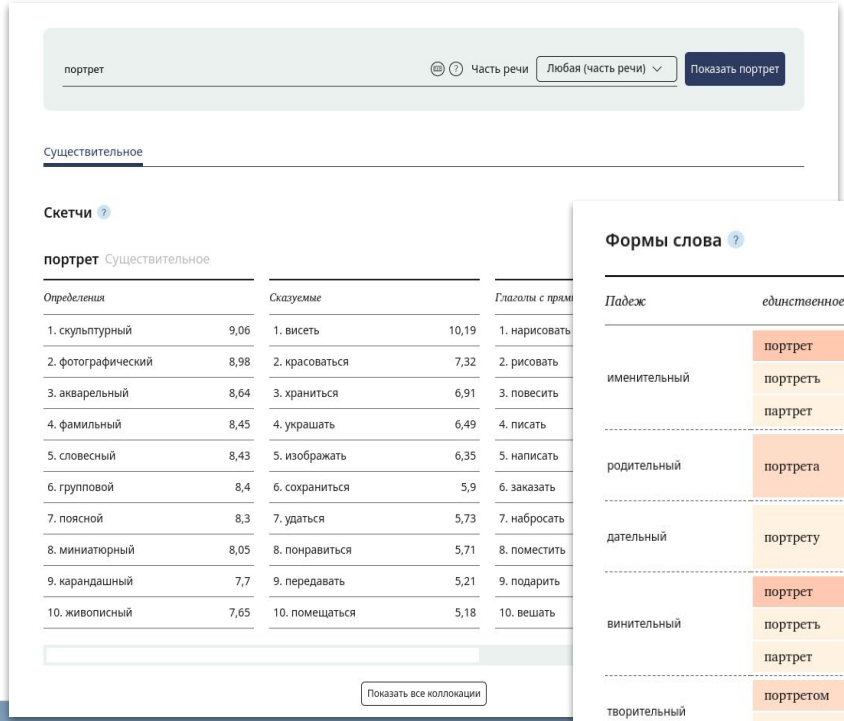
Употребимость синонимов

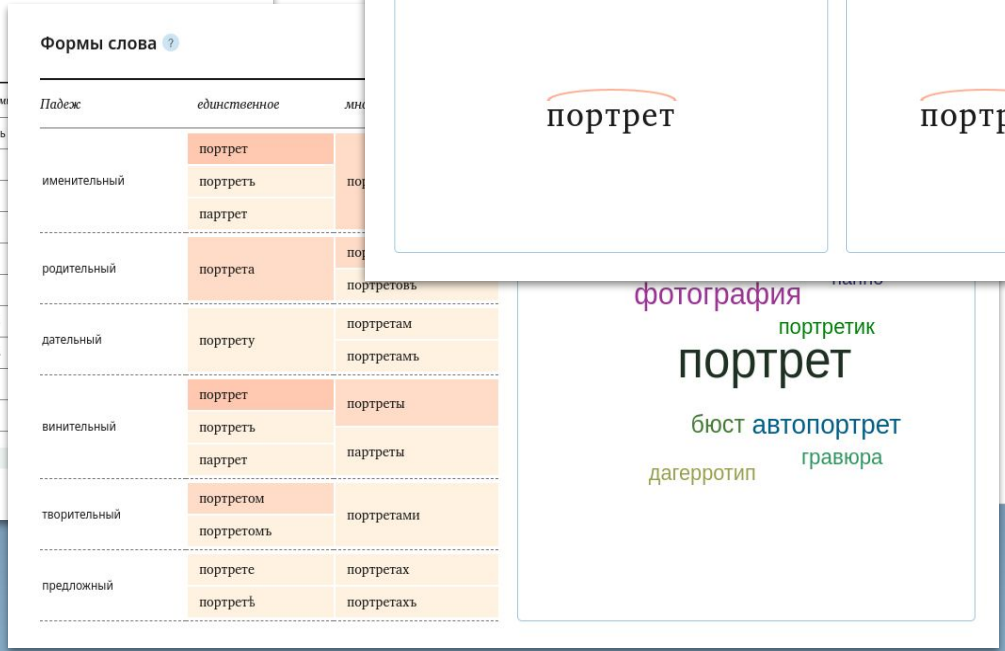
Запросы Графики

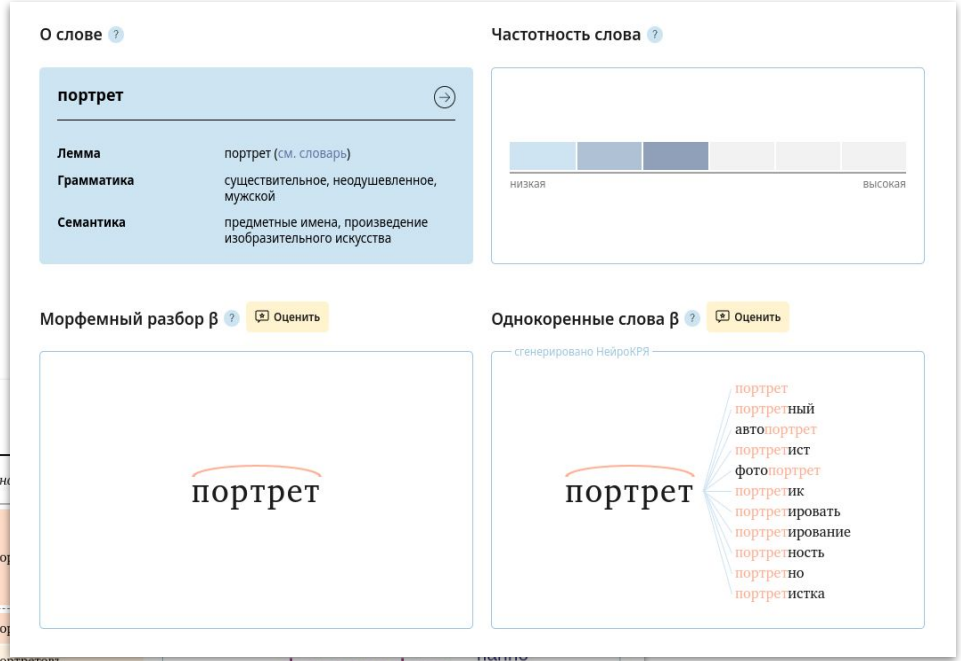
	Запрос 1	*Запрос 2*	*Запрос 3*
	Запрос		
Поиск учитывает:	предпочтительные разборы	предпочтительные разборы	предпочтительные разборы
При поиске:	соседние слова могут совпадать	соседние слова могут совпадать	соседние слова могут совпадать
	Слово 1		
Лемма	смартфон	мобильный	сотовый
	Слово 2		
Лемма	-	телефон	телефон
Расстояние	-	на расстоянии от **1** до **1** от **Слова 1**	на расстоянии от **1** до **1** от **Слова 1**

Портрет слова

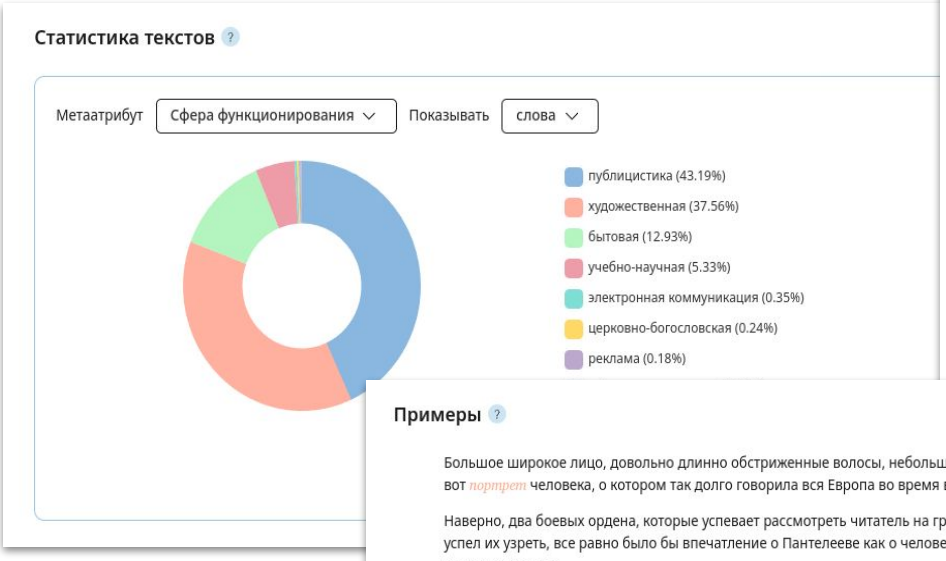
Портрет слова

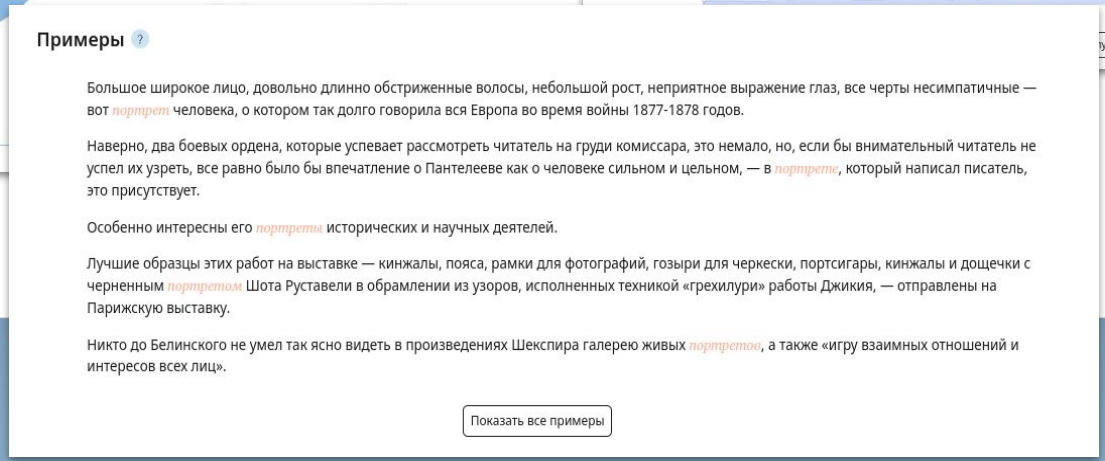


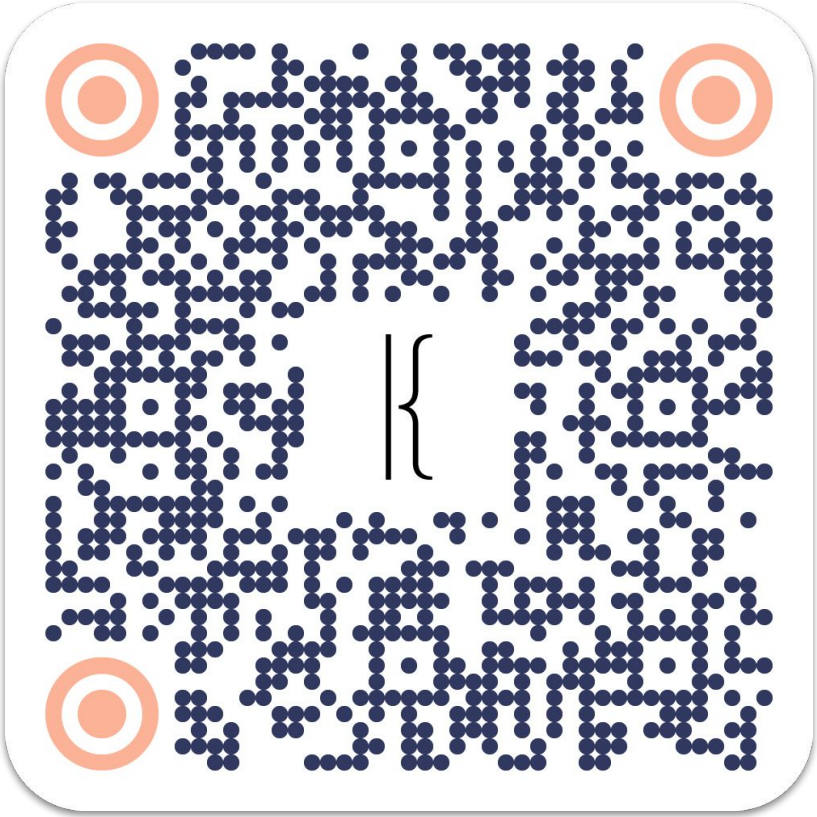


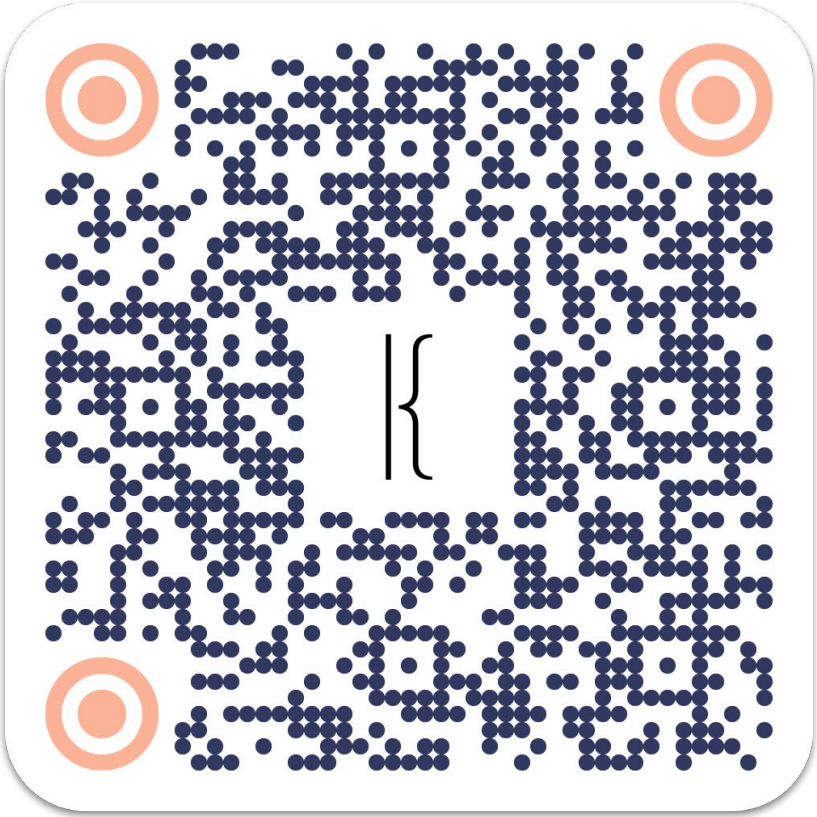


Портрет слова









Спасибо за внимание!

Дмитрий Морозов

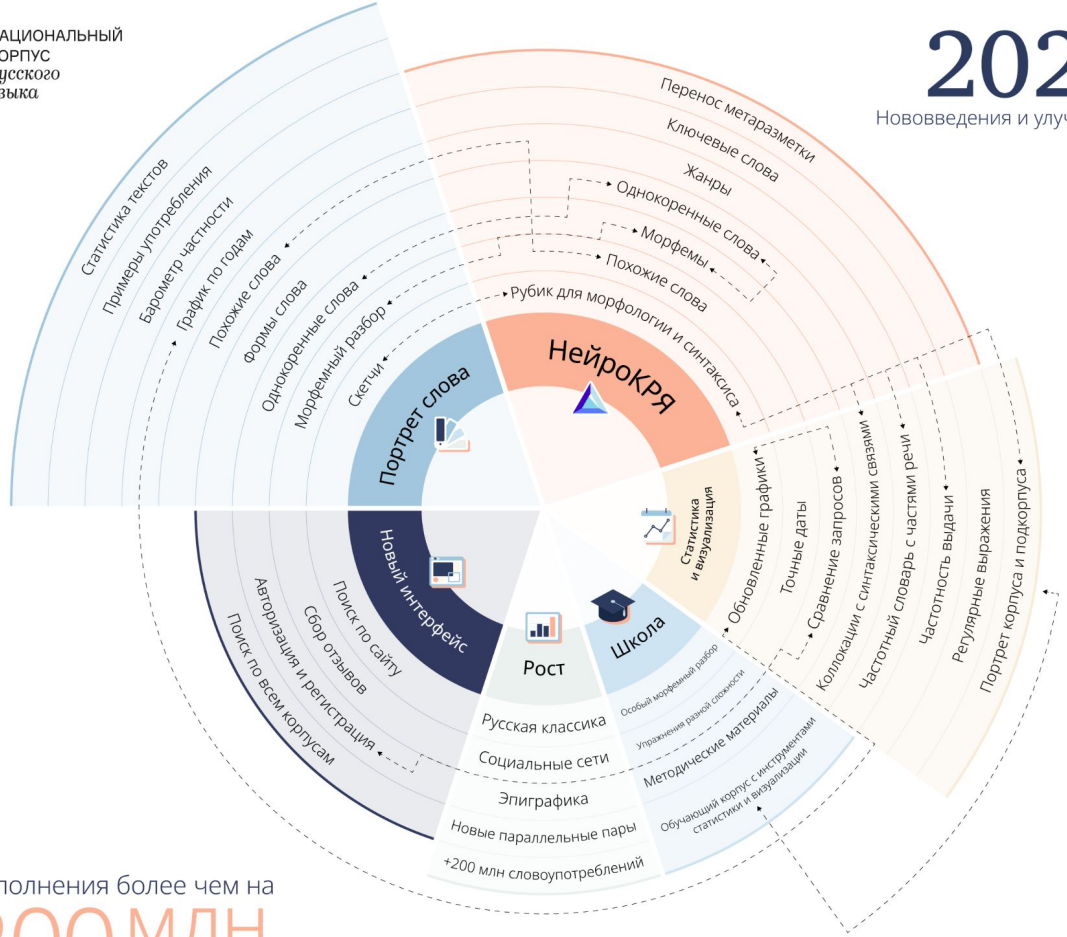
Лаборатория прикладных цифровых технологий ММЦ НГУ
Национальный корпус русского языка

morozowdm@gmail.com | t.me/morozowdm

НАЦИОНАЛЬНЫЙ КОРПУС *русского* языка

2023

Нововведения и улучшения



Пополнения более чем на

200 МЛН словоупотреблений

ИРЯ ИППИ РАН Яндекс