

Филологические базы данных

daniel.tiskin@gmail.com · 9 февраля 2022 г.

1	Компьютерная лексикография и морфология	1
	Грамматические словари	1
	Викисловарь	2
2	Лингвистика в Интернете	3
	Поисковые системы	3
	Социальные сети	4
3	Google Ngram Viewer	5
	Понятие о корпусах	5
	Google Ngram Viewer	5
4	Национальный корпус русского языка	9
	Введение	9
	Лексико-грамматический поиск	12
	Подкорпусы основного корпуса	16
	Другие корпуса НКРЯ	16
5	Интернет-корпусы. CQL	19
	Агапеа	19
	Корпусы Университета Лидса	19
	Создание корпуса	20
6	Обработка данных корпуса	20
7	Базы данных научной информации по лингвистике	25
	О языковом разнообразии	25
	О языковых явлениях	27
8	Базы источников для исследователей	29
	Поиск литературы	29
	Ресурсы СПбГУ	30
	Старые книги	32

1 Компьютерная лексикография и морфология

Грамматические словари В них указано, как изменяется то или иное слово. Самый известный — «Грамматический словарь русского языка» А. А. Зализняка; там исчислены все словоизменительные типы (склонения, спряжения со всеми их вариантами).

Такую информацию можно использовать в компьютерном анализе и порождении текста: редкие типы склонения и спряжения как правило непродуктивны, так что новые (в т. ч. неизвестные компьютеру) слова можно автоматически склонять/спрягать, приписав им одну из уже данных меток. Задавать формы списком не нужно.

- Словарь **инверсионный** — выстраивает слова в алфавитном порядке их концов
- В один «грамматический разряд» попадают слова, изменяющиеся по одному и тому же набору грамматических категорий (например, сравнительная степень, «выпадающая» из типично именной парадигмы прилагательного, выносится в отдельный разряд; местоимение разбросано по разным разрядам)

- Ударение, а точнее рисунок его сдвигов при изменении слова, рассматривается как часть словоизменения и указывается в индексе словоизменительного типа
- Чередования в основе, такие как *лев* — *льва*, тоже указываются
- Все типы проиллюстрированы образцами, так что по индексу типа можно просклонять/проспрягать слово, повторяя словоизменение образца

Несколько (смежных) примеров из словаря:

вы́сылать нсв 1а ⑥(вы́слать) ← несов. вид, спрягается как *делать*, ударение как в *делать*; парный сов. вида — типа 6
ма́ть жо 8е Δ *В. ед.* = ж. р., одуш., склоняется как *грудь*, ударение как в *зуб*; имеет нерегулярность, указанную после Δ
И. ед.; прочие формы ← от *ма́ть* жо 8е
дрема́ть нсв нп 6с ← несов. вид, непереходный (⇒ нет страд. прич.), спрягается как *трепать*, ударение как в *писать*
задрема́ть св нп 6с

Сейчас словарь доступен [в Сети](#) (строка для поиска находится сверху). Обозначения объясняются в [руководстве](#) по использованию. При поиске можно вводить только начало слова, а можно заменить начало слова на *.

Задание 1. Найдите в словаре Зализняка все слова, оканчивающиеся на ...*кура*. —

Викисловарь Информацию о том, как склоняется или спрягается русское слово, проще искать в [Викисловаре](#) — словаре, который может редактировать любой. Окошко для поиска находится справа сверху. Сведения о склонении и спряжении в этом словаре обычно берутся из словаря Зализняка, а примеры употребления (там, где они есть) — из Национального корпуса русского языка, о котором см. ниже.

Задание 2. Найдите в Викисловаре статью для слова *лес*. В разделе **Морфологические и синтаксические свойства** найдите таблицу склонения. Откуда два лишних падежа? —

Морфологические и синтаксические свойства [\[править \]](#)

бе́рег

Существительное, неодушевлённое, мужской род, 2-е склонение (тип склонения 3с(1) по классификации А. А. Зализняка). В сочетаниях типа *на́ берег* ударение может падать на предлог; слово «*берег*» при этом превращается в *клитику*.

Составляет **омоформы** с глаголом *беречь*.

Корень: **-берег-** [\[Тихонов, 1996\]](#).

падеж	ед. ч.	мн. ч.
Им.	бе́рег	берега́
Р.	бе́рега	берего́в
Д.	бе́регу	берега́м
В.	бе́рег	берега́
Тв.	бе́регом	берега́ми
Пр.	бе́реге	берега́х
М.	берегу́	—

Этот фрагмент страницы Викисловаря порождался (на последнем этапе) кодом

```
=== Морфологические и синтаксические свойства ===
{{сущ ru m ina 3с(1)
|основа=бе́рег
|основа1=берег
|слоги={{по слогам|бе́|рег}}
|M=берегу́
|клитика="на́ берег"
}}
Составляет [[Приложение:Омоформы русского языка/6|омоформы]] с глаголом [[беречь]].
{{морфо-ru|берег|и=т}}
```

Фрагмент кода используемого шаблона `сущ ru m ina 3c(1)` выглядит как

```
|nom-sg={{основа}}  
|nom-pl={{основа1}}á  
|gen-sg={{основа}}a  
|gen-pl={{основа1}}óv  
|dat-sg={{основа}}y  
|dat-pl={{основа1}}ám
```

2 Лингвистика в Интернете

Поисковые системы Поисковые системы, такие как Google и «Яндекс», предназначены для поиска информации, максимально интересной пользователю, а не для беспристрастной выдачи всех открытых данных. Но если искать особым образом, то можно выяснить, насколько **много** употреблений определённого типа, а если искать по отдельности разные употребления, то можно выяснить, насколько **часто** относительно других употребляется то или иное слово, выражение и т. п. Тем не менее, результат всё равно зависит от технологии поиска, которую мы не контролируем.

Язык запросов «Яндекса»; язык запросов Google.

"слово" найти точно то, что мы хотим (слово в определённой орфографии, конкретное словосочетание), например *гвардии капитан*

Задание 3. Сравните в Google или «Яндексе» результаты поиска по фразе в третьей половине дня и по фразе "в третьей половине дня".

-слово поиск сайтов, где **нет** этого слова

Задание 4. Сравните в Google или «Яндексе» результаты поиска для "чемпионат мира" и для "чемпионат мира" -футболу.

!слово поиск сайтов, где **обязательно** есть это слово («Яндекс»: +слово)

site:spbu.ru поиск только по сайту СПбГУ

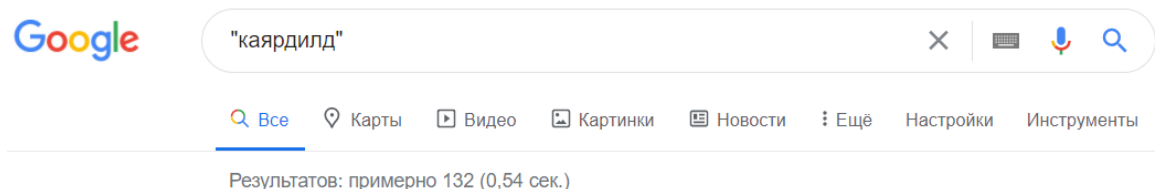
after:2012-07-30 в Google: поиск страниц, написанных после 30 июля 2012 года

before:2017-04-21 в Google: поиск страниц, написанных до 21 апреля 2017 года

Задание 5. Найдите на сайте СПбГУ все употребления сочетания *Барак Обама*, написанные до 30 июня 2016 года, а отдельно — написанные после этого дня.

Найдите на сайте СПбГУ все употребления сочетания *Дональд Трамп*, написанные до 30 июня 2016 года, а отдельно — написанные после этого дня.

Всё это можно использовать, чтобы выяснить, как люди пишут в разных случаях. Когда результаты найдены, Google сообщает, сколько он нашёл:



Но это число может быть неверным. Чтобы получить приближение более высокого качества, нужно зайти на последнюю страницу результатов и посмотреть там. Для серьёзных исследований нужно просматривать каждый результат: вдруг вы ищите сочетание слов, а вам показывают сайт, где начало фразы в одном предложении, а конец в другом?

26 апр. 2020 г. - "Если на Марсе была жизнь 4 миллиарда лет назад, то она может существовать там и **сегодня**. **На Марсе** не было ничего такого, что ...

Задание 6. Определите, сколько раз встречается на сайте МГУ (msu.ru) имя *А. А. Реформатский*, а сколько — *Л. Р. Зиндер*. Прodelайте то же для сайта СПбГУ.

Если мы хотим понять, как говорят чаще, нужно понимать, что с чем сравнивается.

Задание 7. Определите, на каком сайте: СПбГУ или МГУ — слово *ген*тивный* ‘стоящий в форме родительного падежа’ чаще пишется через *е*. —

Чтобы это сделать, нужно найти число x написаний *генитивный* на сайте СПбГУ и число y написаний *генетивный* на сайте СПбГУ; найти долю написаний через e как

$$z = \frac{y}{x + y};$$

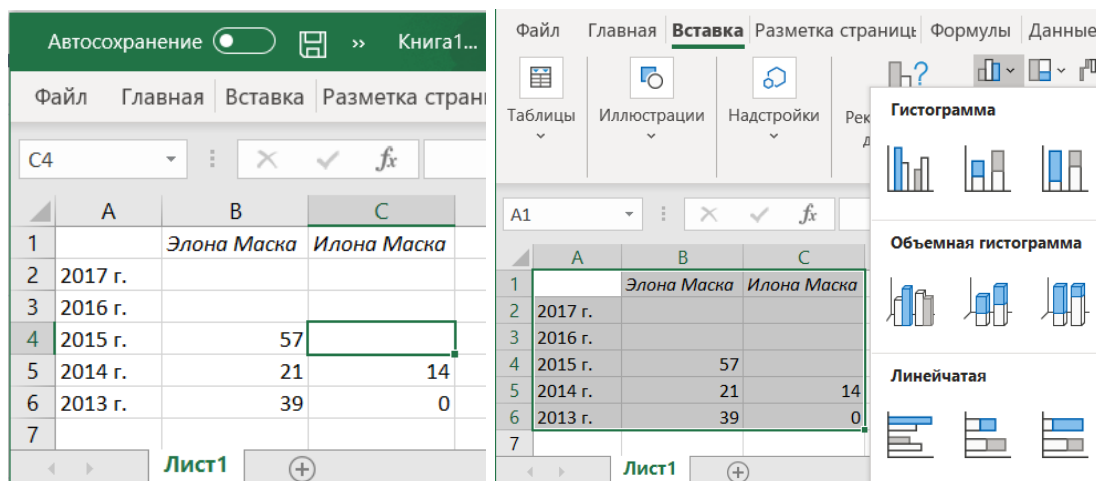
затем найти число x' написаний *генитивный* на сайте МГУ и число y' написаний *генетивный* на сайте МГУ; найти долю написаний через e как

$$z' = \frac{y'}{x' + u'};$$

а затем сравнить z и z' .

Социальные сети С некоторой осторожностью можно использовать для лингвистического исследования социальные сети. Например, можно сравнивать разные написания одного слова, пользуясь [расширенным поиском Twitter](#). Если вам нужна фраза целиком, пользуйтесь полем **Точная фраза**. Пусть мы хотим узнать, когда чаще писали *Элон Маск*, а когда *Илон Маск*. Для примера рассмотрим форму родительного/винительного *Э/Илона Маска* в период с 1 по 31 июля в разные годы. Сначала вводим точную фразу **Элона Маска**, а затем прокручиваем вниз и в разделе **Даты** выбираем нужный период.

Нажимаем кнопку **Поиск**. На полученной странице выбираем **Последнее**. Теперь можно считать результаты. Чтобы сравнить написания за конкретный год, делаем новый поиск, где ищем **Илона Маска** за **тот же** период времени.



По мере подсчёта будем вводить данные в электронную таблицу. Когда все данные введены в таблицу, можно выделить всю таблицу и построить диаграмму, которая наглядно показывает соотношение чисел за каждый год.

В реальном исследовании бывает полезно не считать за разные случаи одно и то же предложение на разных сайтах: при цитировании пишущий не выбирает форму.

3 Google Ngram Viewer

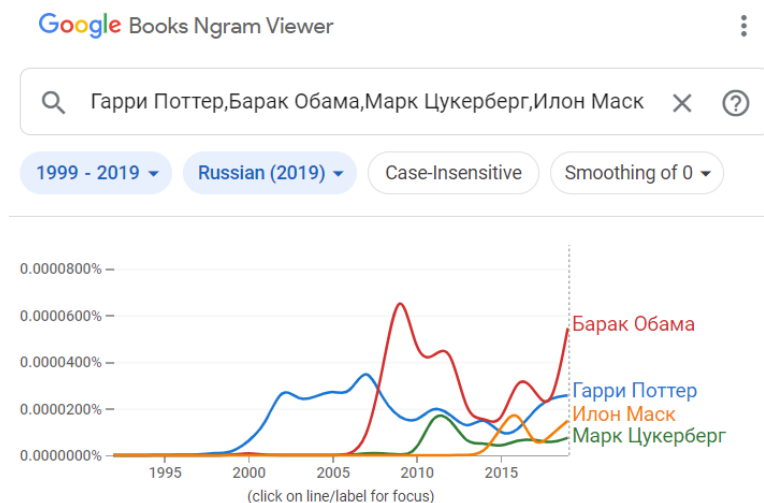
Понятие о корпусах Если вас спросить, насколько чаще говорят *со мной*, чем *со мною*, вы сразу не скажете. Так же и если вас спросить, когда перестали говорить *аэроплан* и стали говорить только *самолёт* или что чаще: *Nabokov's novels* или *novels by Nabokov*. На такие вопросы **можно ответить**, если у вас есть **корпус** — большая коллекция текстов, написанных разными людьми в разное время. Разное время написания важно для вопросов вроде «Когда стали говорить?..», большой размер — для точности подсчётов. Большая коллекция текстов в наше время — это как минимум миллионы слов. Например, **Национальный корпус русского языка** (НКРЯ) имеет объём около 320 млн слов (в основной части), а в крупнейшем из корпусов русского языка из семейства **Aranea** почти 20 млрд слов. Такие объёмы данных не просмотреть вручную и сложно обрабатывать с помощью обычного поиска. Поэтому большие корпуса иногда делают в сотрудничестве с компаниями, занимающимися поиском в Интернете. Например, НКРЯ обслуживается «Яндексом», а Google сделал свои корпуса сам.

С точки зрения состава корпус может представлять конкретную тему, жанр, тип коммуникации или регион либо комбинацию этих параметров (например, только устная речь людей из определённого города, как в «**Одном речевом дне**»). Может быть и сбалансированный корпус, отражающий язык в совокупности регистров и периодов, как в НКРЯ. Но некоторые ценные результаты даёт простой поиск по большому количеству книг. Так устроен ресурс Google Ngram Viewer, о котором мы будем говорить сначала.

Google Ngram Viewer Этот ресурс основан на книгах, оцифрованных и размещённых в **Google Books** (где доступны для чтения не все из них). Выбор книг и качество оцифровки не контролируются нами, поэтому **не всем подсчётам можно доверять**. Особенно это касается лексики (много советских книг, где часто встречаются слова, которыми мы уже не пользуемся) и отчасти морфологии и орфографии.

Как пользоваться поиском, описано в **руководстве**.

Ngram Viewer показывает, какой была частота употребления данного слова или выражения (в процентах от всех слов) в данный год. Чтобы построить график, нужно ввести интересные нас слова или выражения **через запятую**.

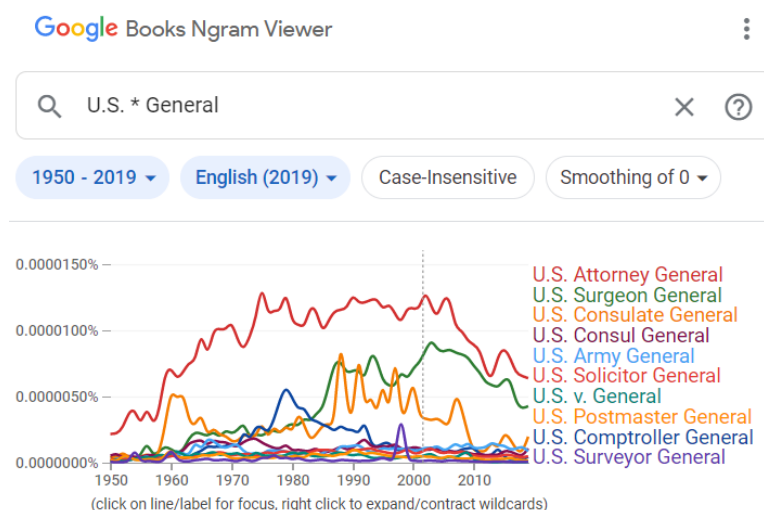


Под полем поиска есть кнопка `1800 - 2019`, где можно ввести период времени, который интересует нас. Кнопка `English 2019` позволяет выбрать интересующий нас корпус. Если нас интересует русский язык, выбираем `Russian 2019`. Кнопка `Case-Insensitive` позволяет выбрать, вместе или по отдельности считаются *ЗАГЛАВНЫЕ* и *строчные* буквы. На кнопке `Smoothing` лучше выбрать 0, чтобы видеть настоящие графики; другое число даст искажённый график. Когда мы введём слова и нажмём Enter, будет построен график (частота в процентах по вертикальной оси).

Задание 8. Повторите поиск, показанный на рисунке выше. ─

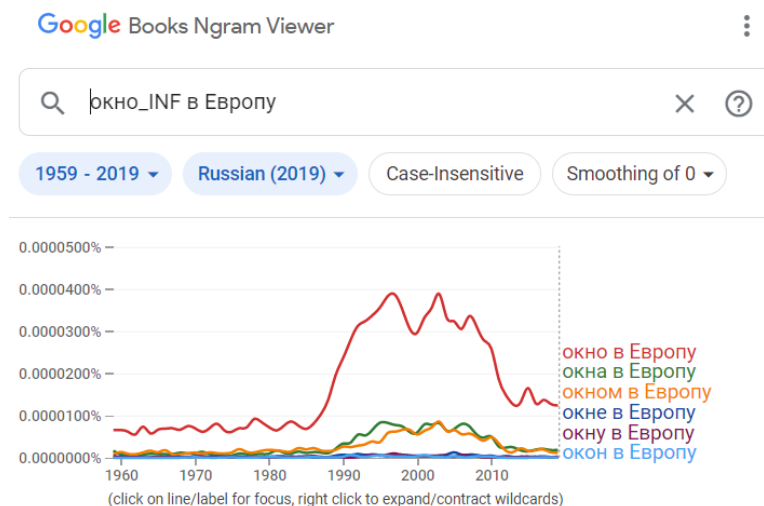
Задание 9. Выбрав корпус английского языка за период с 1819 по 2019 год, определите, в какой момент стали писать *tomorrow* чаще, чем *to-morrow* (или *today* чаще, чем *to-day*). Используйте сглаживание (smoothing) 0. ─

Можно искать сразу несколько сочетаний слов, имеющих общую часть и различающихся одним словом. Для этого переменную часть обозначают знаком *. Мы получим 10 самых частых результатов за выбранный период. Если нажать **правой** кнопкой на название любого из вариантов, получится общий график.



Задание 10. Повторите поиск, показанный на рисунке выше. Получите общий график нажатием правой кнопки мыши. ─

Задание 11. Выбрав корпус английского языка и период с 1998 по 2008 год, определите, какие слова чаще всего встречались после сочетания *Harry Potter and* и когда какое из них начинало употребляться часто. ─

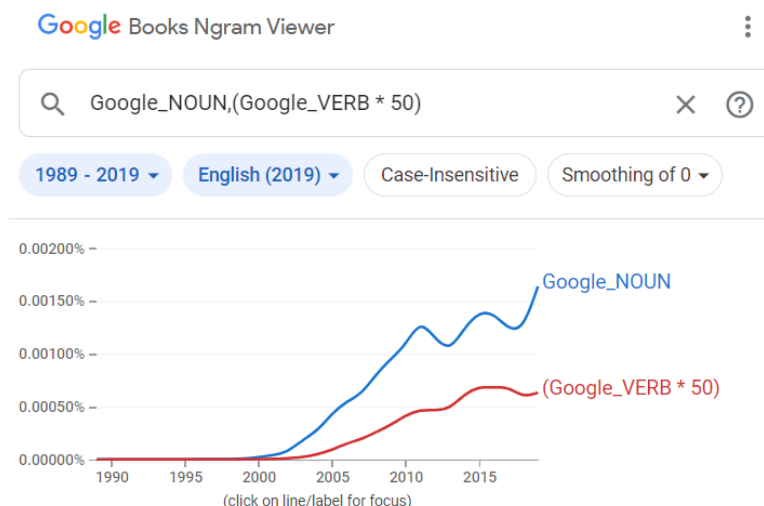


Что делать, если вы хотите посмотреть употребление слова во всех его формах вместе? Если искать **заяц**, не будут найдены формы *зайца*, *зайцу*, *зайцы* и т. д. Чтобы искать их вместе, нужно написать в строке поиска **заяц_INF** (видимо, от *inflection* ‘изменение слова по формам’). Но такое слово может быть только одно в запросе.

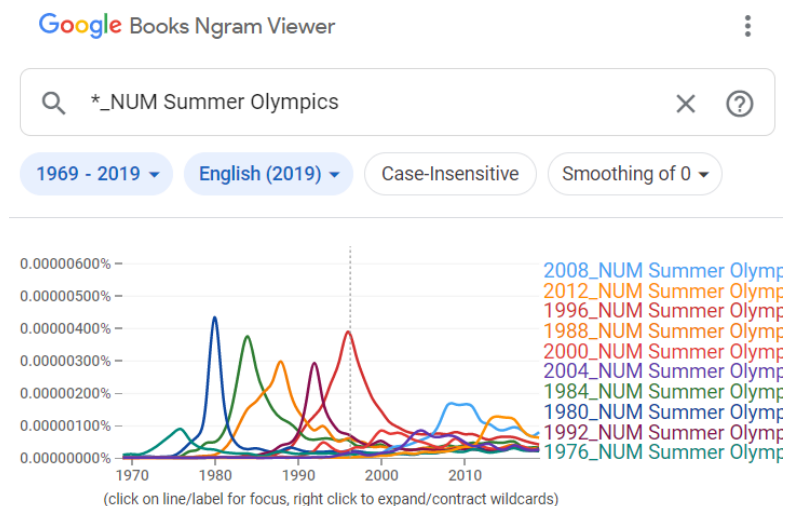
Задание 12. Повторите поиск, показанный на рисунке выше. Получите общий график нажатием правой кнопки мыши.

Задание 13. Выбрав корпус русского языка и период с 1899 по 2019 год, определите, как менялась частота употребления **всех форм** слова **аэроплан вместе**. Постройте на том же графике кривую для слова **самолёт** во всех его формах.

Кроме того, есть формы слов, которые можно отнести к разным частям речи: *мой* — это и местоимение в начальной форме, и повелительное наклонение от *мыть*; *сыр* — и существительное в начальной форме, и краткая форма от *сырой*. Чтобы найти форму в функции конкретной части речи, нужно написать обозначение этой части речи (список обозначений можно найти в руководстве).

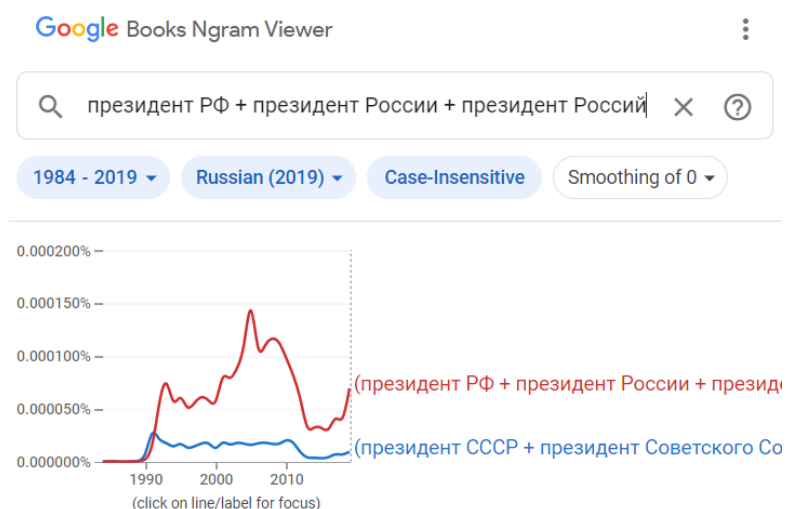


На рисунке выше один из графиков увеличен в 50 раз, чтобы его было лучше видно. Можно искать **разные** сочетания, в которых на определённом месте какое-то слово интересующей нас части речи:

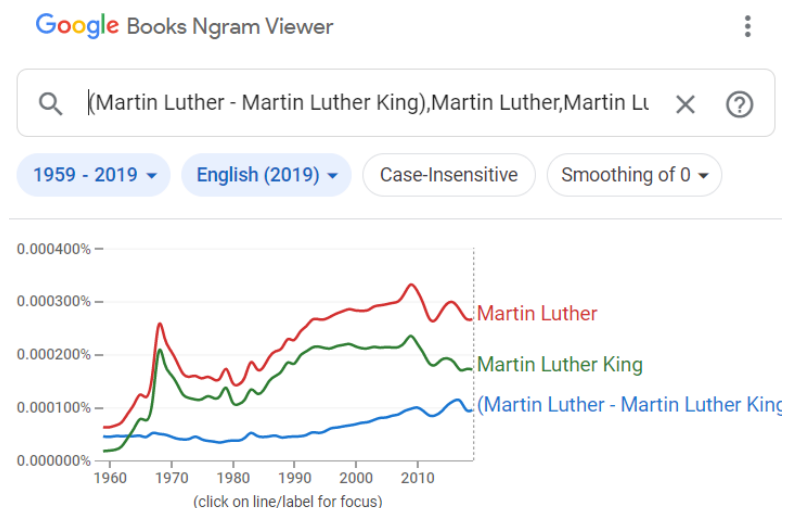


Задание 14. Выбрав корпус русского языка и период с 1919 по 2019 год, определите, как менялись наиболее частотные прилагательные (ADJ) в сочетании *ветеран ... войны*. —

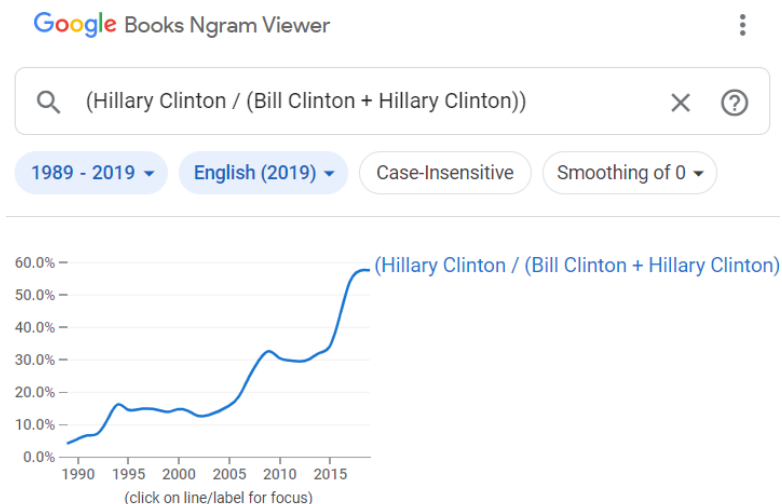
Графики можно складывать, вычитать и делить. Узнаем, как менялась частота сочетаний *президент РФ*, *президент России* и *президент Российской Федерации* вместе и как менялась частота сочетаний *президент СССР* и *президент Советского Союза* вместе:



Вычитание полезно, когда нужно что-то исключить из подсчёта, например выяснить, как менялась частота употребления имени *Martin Luther* без учёта случаев, когда это на самом деле *Martin Luther King*:



Деление нужно, чтобы найти долю интересующих нас употреблений от всех похожих:



4 Национальный корпус русского языка

Введение Национальный корпус русского языка (НКРЯ) — один из основных исследовательских ресурсов для тех, кто занимается современным состоянием русского языка и его историей на протяжении последних нескольких веков. Он же и один из самых понятных, так что его легко освоить на базе школьного курса. Это не касается синтаксического корпуса (см. ниже), в отличие от понятной синтаксической разметки в [Хельсинкском аннотированном корпусе](#) (ХАНКО).

Задание 15. Найдите в ХАНКО все вхождения существительных, чья начальная форма заканчивается на *-ка*, в творительном падеже в составе односоставного придаточного. Объясните двойной синтаксический разбор в первом найденном примере. ─

Сайт НКРЯ — <http://ruscorpora.ru>, страница поиска доступна там по ссылке [поиск в корпусе](#) слева. По этой ссылке находится следующий интерфейс:

НАЦИОНАЛЬНЫЙ КОРПУС
РУССКОГО
ЯЗЫКА

главная Основная корпус инструкция задать подкорпус English

основной Поиск точных форм 1 A B B 7

– корпус Слово или фраза

– биграммы 2

– триграммы искать очистить

– 4-граммы

– 5-граммы

синтаксический Лексико-грамматический поиск 3

газетный Слово ? A B B Грамм. признаки ? выбрать Семант. признаки ? выбрать

параллельный Доп. признаки ? выбрать 1-е знач. [x] др. знач. [x] фильтр 1 [] фильтр 2 []

обучающий Расстояние: от 1 до 1 6

диалектный 8

поэтический 5 признаки ? выбрать 1-е знач. [x] др. знач. [x] фильтр 1 [] фильтр 2 [] 4

устный

акцентологический

мультимедийный

мультипарк

исторический

- ① «Поиск точных форм» — сюда можно ввести слово или сочетание слов, чтобы найти все случаи, когда это слово или сочетание употреблено в текстах, данных в корпусе.
- ② Эту кнопку нужно нажать, если мы делаем поиск точных форм.
- ③ «Лексико-грамматический поиск» — здесь в каждое поле под словом **Слово** можно написать только одно слово! Оно должно быть в форме, в которой его пишут в словаре, например *нести*, а не *несу* или *несущий*; *коза*, а не *kozy* или *козам*. Если нужно несколько слов, второе пишется под следующим словом **Слово**...
- ④ ...А чтобы было третье, нужно около второго нажать кнопку справа. Чтобы убрать лишнее слово, нужно нажать кнопку .
- ⑤ Эту кнопку нужно нажать, если мы делаем лексико-грамматический поиск.
- ⑥ Расстояние между словами в лексико-грамматическом поиске: от до означает, что мы ищем соседние слова; от до — либо соседние, либо такие, между которыми ещё одно слово; и т. д.
- ⑦ «Задать подкорпус» — меню выбора частей корпуса, в которых мы хотим искать. Если туда не заходить, будем искать по всему корпусу.
- ⑧ Разные другие корпуса в составе НКРЯ. Например, газетный устроен так же, как и наш «основной», но тексты там взяты из газет, а не из книг, научных журналов или Интернета, как многие в основном корпусе. В историческом корпусе собраны древнерусские и вообще старые тексты, где другие орфография, грамматика и нужно учиться искать отдельно.

Напишем в строке поиска точных форм прожорливое брюшко и нажмём .

[перейти на страницу поиска](#) [выбрать подкорпус](#) [версия с ударениями](#) [настройки](#) [формат KWIC](#) [English](#)

8 10 9

Результаты поиска в основном корпусе

Объем всего корпуса: 124 565 документов, 321 712 061 слово. 1

"прожорливое" на расстоянии 1 от "брюшко" 2

Найдено 1 документ, 1 вхождение. 3

[Распределение по годам](#) [Статистика](#)

Поискать в других корпусах: [акцентологическом](#), [газетном](#), [диалектном](#), [мультимедийном](#), [обучающем](#), [параллельном](#), [поэтическом](#), [устном](#).

Страницы: 1 4 5 6

1. Николай Носов. Приключения Незнайки и его друзей (1953-1954) [омонимия не снята] [Все примеры](#)

7 Но вот пришла лягушка, **Прожорливое брюшко**, И съела кузнеца, И съела кузнеца. [Николай Носов. Приключения Незнайки и его друзей (1953-1954)] [омонимия не снята] [←...→](#)

Страницы: 1

Поискать в других корпусах: [акцентологическом](#), [газетном](#), [диалектном](#), [мультимедийном](#), [обучающем](#), [параллельном](#), [поэтическом](#), [устном](#).

Скачать несколько первых результатов выдачи в формате [Excel](#), [CSV](#).

- ① Размер всего корпуса. «Документы» — это отдельные тексты, из которых он состоит.
- ② Описание запроса — параметры, по которым мы искали тексты.
- ③ Количество найденных текстов и «вхождений» — случаев употребления того, что мы искали. Бывает, что в одном тексте несколько случаев.
- ④ Страницы с найденными примерами (**конкорданс** — поисковая выдача с выделенным материалом, релевантным запросу). Здесь всего одна; если найдено больше, здесь будут ссылки на другие страницы.

- ⑤ Название текста. Если нажать на него, появится карточка с описанием текста.
- ⑥ Пометка об омонимии: [омонимия не снята] означает, что в этом тексте форма *стол* будет описана и как форма именительного, и как форма винительного падежа; форма *мела* — и как форма родительного падежа существительного *мел*, и как форма прошедшего времени глагола *мести*; и т. д. Если написано [омонимия снята], то в этом тексте для каждого такого случая выбран правильный разбор.
- ⑦ Отдельный пример употребления. В скобках в конце написано, откуда он взят. Пример можно скопировать и вставить в работу по лингвистике, которую вы пишете; название в скобках поможет читателю понять, откуда пример. Сейчас по данным корпуса пишут целую [грамматику](#) русского языка.
- ⑧ Если нажать сюда, в примерах появятся ударения.
- ⑨ Если нажать сюда, результаты будут выстроены по слову, которое вы искали:

Илья Лагутенко — это **Бильбо** Бэггинс и Гарри Поттер в [←...→](#)
 в роли старого хоббита-скупердя **Бильбо** . [←...→](#) Мария Васильева. Хоббиты пришли // «Вечерняя Москва», 2002.02.07
 неплохо меня отдал. — Ну что, **Бильбо** Бэггинс, изучил город? Голос был [←...→](#)
 сравнимо с посиделками гномов у **Бильбо** накануне их путешествия. [←...→](#)
 Взломщик **Бильбо** Бэггинс стал простым хакером, а [←...→](#)
 Что? Некто по имени **Бильбо** , говорите вы? Коротенький, толстенький, с [←...→](#)
 Малая планета № 2991. Астероид **Бильбо** ... Однажды известный маг Гэндальф зашел [←...→](#)
 одного спутника, почтенного хоббита мистера **Бильбо** Бэггинса, эскайра. [←...→](#)
 И в пути нашел **Бильбо** золотое колечко, которое долгие века [←...→](#)
 одного из гномов — спутников хоббита **Бильбо** Бэггинса. [←...→](#)
 Есть, правда, культурная тимофеевка сорта « **Бильбо** », но это отнюдь не таксон [←...→](#)
 Хоббит **Бильбо** Бэггинс, кстати, тоже ведь холостяк [←...→](#)
 Хоббит **Бильбо** Бэггинс просто случайно забрел на [←...→](#)
 всем, что умудрился напутать этот **Бильбо** . [←...→](#)
 по природе стал обыкновенный хоббит **Бильбо** Бэггинс. [←...→](#)
 удивляло друзей и родичей в **Бильбо** , а потом и в его [←...→](#)

- ⑩ Если нажать сюда, появится окно, в котором можно выбрать, сколько примеров на каждой странице будет показывать корпус и в каком порядке он будет их показывать.

Упорядочить:

по умолчанию

по умолчанию

случайно

по автору

по дате создания

по дате создания в обратном порядке

по дате рождения автора

по дате рождения автора в обратном порядке

5

слов в контексте KWIC-выдачи

50

документов на странице

10

примеров в документе

50

примеров на странице

Применить

Восстановить исходные значения

Если ничего не делать, корпус сначала будет показывать примеры [омонимия снята] (новые тексты выше, старые ниже), а потом примеры [омонимия не снята] (тоже новые выше, а старые ниже).

Ссылка [Распределение по годам](#) выдаёт график, на котором показано, насколько часто то слово, которое мы ищем, встречается в текстах за тот или иной год (как в Google Ngram Viewer). Но в НКРЯ не очень много текстов, особенно за давние годы, поэтому получившимся результатам нельзя полностью верить. Кроме того, в поиске точных форм ищется только одна форма слова, а употребление других не будет показано на графике. Доступно и [отдельно](#); работает и для нескольких слов или сочетаний.

Задание 16. Сравните динамику употребления сочетаний *генералиссимус Сталин*, *генерал Эйзенхауэр*, *команданте Че*, *субкоманданте Маркос* в 1938–2018 гг. —

Задание 17 (студенты МКК, 2017). Чтобы убедиться в опасности графиков (без конкорданса), посмотрите при значительном сглаживании, как распределяются по годам слова *компьютер*, *квартильный*, *орк*, *дракон*. Попробуйте предложить объяснение и проверьте свою гипотезу, меняя сглаживание и обращаясь к результатам поиска. —

Руководство по поиску в корпусе (рекомендую разделы 4.1, 4.4–4.8, 4.10–4.12). Краткие **видеоуроки** О. Н. Ляшевской. См. также **примеры запросов**.

Помимо грамматических задач (есть целая **грамматика**), корпуса полезны в лексикографии; см. **словари** на основе НКРЯ, из которых чаще всего упоминается **частотный** (сравнительно старый, так что использует не весь объём современного подкорпуса).

Лексико-грамматический поиск Полностью использовать возможности корпуса можно, если вы ищете слова и сочетания слов с определёнными грамматическими признаками, например числом, падежом, наклонением, лицом и т. д. Тогда вы сможете находить **любое** слово с такими признаками, а не только какое-то одно, как бывает, если записать само слово в поле **Слово** лексико-грамматического поиска.

Чтобы выбрать эти признаки, можно нажать на ссылку **Грамм. признаки ? выбрать**. Тогда выскочит меню, где все признаки перечислены на одном экране:

Часть речи <input type="checkbox"/> существительное <input type="checkbox"/> прилагательное <input type="checkbox"/> числительное <input type="checkbox"/> числ-прил <input type="checkbox"/> глагол <input type="checkbox"/> наречие <input type="checkbox"/> предикатив <input type="checkbox"/> вводное слово <input type="checkbox"/> мест-сущ <input type="checkbox"/> мест-прил <input type="checkbox"/> мест-предикатив <input type="checkbox"/> местоименное наречие <input type="checkbox"/> предлог <input type="checkbox"/> союз <input type="checkbox"/> частица <input type="checkbox"/> междометие	Падеж <input type="checkbox"/> именительный <input type="checkbox"/> звательный* <input type="checkbox"/> родительный <input type="checkbox"/> родительный 2 <input type="checkbox"/> дательный <input type="checkbox"/> винительный <input type="checkbox"/> винительный 2* <input type="checkbox"/> творительный <input type="checkbox"/> предложный <input type="checkbox"/> предложный 2 <input type="checkbox"/> счётная форма	Наклонение / Форма <input type="checkbox"/> изъявительное <input type="checkbox"/> повелительное <input type="checkbox"/> повелительное 2 <input type="checkbox"/> инфинитив <input type="checkbox"/> причастие <input type="checkbox"/> деепричастие	Степень / Краткость <input type="checkbox"/> сравнительная <input type="checkbox"/> сравнительная 2 <input type="checkbox"/> превосходная <input type="checkbox"/> полная форма <input type="checkbox"/> краткая форма
	Число <input type="checkbox"/> единственное <input type="checkbox"/> множественное	Лицо <input type="checkbox"/> первое <input type="checkbox"/> второе <input type="checkbox"/> третье	Переходность <input type="checkbox"/> переходный* <input type="checkbox"/> непереходный*
Имена собственные <input type="checkbox"/> фамилия <input type="checkbox"/> имя <input type="checkbox"/> отчество	Род <input type="checkbox"/> мужской <input type="checkbox"/> женский <input type="checkbox"/> средний <input type="checkbox"/> общий*	Залог <input type="checkbox"/> действительный <input type="checkbox"/> страдательный <input type="checkbox"/> медиальный	Прочее <input type="checkbox"/> цифровая запись <input type="checkbox"/> аномальная форма* <input type="checkbox"/> искажённая форма* <input type="checkbox"/> инициал* <input type="checkbox"/> сокращение* <input type="checkbox"/> несклоняемое* <input type="checkbox"/> топоним**
	Одушевленность <input type="checkbox"/> одушевленное <input type="checkbox"/> неодушевленное	Вид <input type="checkbox"/> совершенный <input type="checkbox"/> несовершенный	

* - только в корпусе со снятой омонимией
 ** - только в корпусе с неснятой омонимией

Выбрав нужное, нажмите ОК. В поле **Грамм. признаки** появятся метки, соответствующие тому, что вы выбрали; например, **S, nom** — это существительное в именительном падеже. Поле **Слово** при этом можно не заполнять.

На примере ХАНКО видно, что поиск по морфологическим признакам в корпусе происходит не на ходу, а путём обращения поисковой машины к уже готовой разметке:

```
<a href=showinfo.phtml?lang=r&id=300831 TARGET=sameRight CLASS=nf>улыбается</a>
<a href=showinfo.phtml?lang=r&id=300832 TARGET=sameRight CLASS=nf>и</a>
<a href=showinfo.phtml?lang=r&id=300833 TARGET=sameRight CLASS=nf>машет</a>
<a href=showinfo.phtml?lang=r&id=300834 TARGET=sameRight CLASS=nf>им</a>
<a href=showinfo.phtml?lang=r&id=300835 TARGET=sameRight CLASS=f>пыкой</a>
<a href=showinfo.phtml?lang=r&id=300836 TARGET=sameRight CLASS=nf>.</a>
```

В старой версии НКРЯ эта разметка особенно хорошо видна, поскольку встроена в код страницы напрямую: `ср.`

```
<span class="b-wrd-expl" explain="беседовал|108|28851|0|1550|1|1|1">
беседовал</span>
```

Теоретическая лингвистика говорит об этом примерно следующее. У конкретного употребления конкретной словоформы в корпусе есть пометы, относящиеся к лексеме в целом («постоянные признаки»), и пометы, относящиеся к данной словоформе («непостоянные»). Все эти признаки представляют собой значения различных **грамматических категорий**. Как ни странно, конкретный падеж или конкретное число — это смыслы, а не способы их выражения! Ср. окончания `-□`, `-ов`, `-ей`, `-ых`, выражающие значение GEN.PL в разных типах склонения.

Определение 1. Категорией называется любое наибольшее множество значений, из которых никакие два не могут встречаться вместе в одной позиции.

Скажем, глагольный вид, число или падеж имени. Но не множество {им., род., дат.}.

Определение 2. Грамматической называется категория, обладающая обязательностью, т. е. такая, у которой в каждом случае выражается одно и только одно значение.

Обязательность категории для класса слов может проявиться любым из двух способов:

- каждое слово класса охарактеризовано по данной категории во всей совокупности своих форм — **словоклассифицирующая ГК**
- каждая из форм каждого слова охарактеризована по данной категории, но у разных форм значения категории могут быть различными — **словоизменяющая ГК**

Задание 18. Найдите все местоимения-прилагательные в краткой форме. —

Задание 19. Найдите все формы всех прилагательных, которые не склоняются (грамматический признак 0). —

Если выбрать ☐ существительное и ☐ именительный, то метки **S** и **nom** появятся в поле **Грамм. признаки** через запятую. Но если выбрать ☐ именительный и ☐ дательный, то они появятся в поле **Грамм. признаки** как **(nom|dat)**. НКРЯ знает, что «быть существительным» и «быть в именительном падеже» — это признаки, которые могут быть вместе, а вот «именительный» и «дательный» вместе не бывают, поэтому он воспринимает ваше требование «☐ именительный + ☐ дательный» как «именительный **или** дательный», откуда значок «|». Но можно переделать запись **S,nom** в **(S|nom)**, и тогда корпус будет искать все слова, которые или существительные, или стоят в именительном падеже (включая и те существительные, которые стоят в именительном падеже).

Задание 20. Найдите все слова, помеченные одновременно как «единственное число» и как «множественное число». Для этого сначала из грамматических признаков выберите вместе ☐ единственное и ☐ множественное, а затем поменяйте (...|...) на запятую. —

Задание 21. Найдите все сочетания из двух слов (идущих подряд) в любой форме, из которых первое — *Наум* или *Ноам*, а второе — *Хомский*. Первое делается с помощью конструкции (Ноам|Наум) в поле Слово. ─

Можно комбинировать поле Слово и грамматические признаки.

Задание 22. Найдите все употребления слова *жучок* в творительном или предложном падеже множественного числа. ─

В поле Слово необязательно писать слово целиком. Можно искать слова, начинающиеся определённым образом; например, кот* в поле Слово даст результаты со словами *кот*, *котёнок*, *который* и т. д. Наоборот, *орый в поле Слово даст *который*, *скорый*, *хворый* и т. д., но во всех формах, а не только в тех, которые сами заканчиваются на ...орый. Чтобы найти форму слова, которая сама начинается или заканчивается определённым образом, в поле Слово надо записать всю форму со звёздочкой в кавычках: "*орый" даст только формы, заканчивающийся на ...торый, а "*ория" даст формы *Виктория*, *история* (именительный падеж), *тория* (родительный падеж от *торий*) и т. д.

Задание 23. Найдите все сочетания из двух слов (идущих подряд) в любой форме, из которых первое — *дочь*, а второе — любое слово, оканчивающееся в начальной форме на ...аша или ...юша. ─

Задание 24. Найдите все сочетания из двух слов (идущих подряд) в любой форме, из которых первое — конкретная форма, оканчивающаяся на ...оксидом, а второе — конкретная форма *углерода*. ─

Можно указывать не только признаки, которые вы хотите найти, но и признаки, которых вы хотите избежать (а также и слова в поле Слово, которых вы хотите избежать). Перед каждым таким признаком (в поле Грамм. признаки) или словом (в поле Слово) нужно поставить знак «-»; если таких слов несколько, «-» ставится перед каждым после пробела, а запятые не нужны. Все команды с «-» ставятся строго после всех команд без «-». Например, если мы хотим найти любое слово в конкретной форме, оканчивающейся на ...ыбы, кроме форм слов *рыба* и *глыба*, причём в родительном падеже (gen), но не во множественном числе (pl), то запрос будет выглядеть как

Слово ? <input type="text" value="А Б В"/>	Грамм. признаки ? выбрать
<input *ыбы\"="" -глыба"="" -рыба="" type="text" value="\"/>	<input type="text" value="gen -pl"/>
Доп. признаки ? выбрать	<input type="text"/>

Задание 25. Найдите все сочетания слова *Лутер* в любом падеже, кроме именительного, с любым словом, кроме слова *Кинг*. ─

Задание 26. Найдите все формы всех существительных, начинающихся на *хобот...*, кроме слова *хобот*. ─

Задание 27. Найдите все формы всех прилагательных, оканчивающихся (в начальной форме) на ...конный, кроме слова *конный* и всего, что оканчивается на ...законный. ─

Задание 28. Найдите все употребления глаголов, в конкретной форме оканчивающиеся на ...крякивает. Эту форму нужно заключить в кавычки: "*крякивает". ─

Поле Семантические признаки основано на приписанных лексемам в корпусе меток семантической классификации.

Задание 29. Пользуясь подкорпусом НКРЯ со снятой омонимией, установите, какие слова могут обозначать одновременно людей и еду; людей и инструменты; людей и части растений (кроме слова *мешок*). —

Задание 30. Ограничиваясь законодательными и правовыми документами, определите, какие сверхъестественные существа (кроме *созданий* и *гор(ов)*) там упоминаются. Каузируют ли они какие-либо положения дел? (Сочетания с каузативными глаголами.) —

Дополнительные признаки позволяют ограничить выдачу по какому-то ещё признаку. Так, помета **first** появляется, если выбрать ☐ **в начале предложения**: все найденные слова с этой пометой будут первыми в своих предложениях; **last** (☐ **в конце предложения**) ищет последнее слово в предложении. Можно выбрать, должна ли перед словом быть запятая или другой знак. Как всегда, перед такими пометами можно поставить знак «-»: **-first** — не первое слово в предложении, **-amark** — слово не после знака препинания.

The screenshot shows a web browser window with the URL 'ruscorpora.ru/new/reqflags-filtered.html'. The interface contains several sections of checkboxes for filtering search results:

- Повтор*** (Repeat):
 - ☐ лексемы
 - ☐ части речи
 - ☐ падежа
 - ☐ числа
 - ☐ времени
 - ☐ рода
 - ☐ лица
 - ☐ одушевлённости
- Тип оборота** (Type of phrase):
 - ☐ mw:ADV
 - ☐ mw:ADVPRO
 - ☐ mw:CONJ
 - ☐ mw:PARENTH
 - ☐ mw:PART
 - ☐ mw:PR
 - ☐ mw:PRAEDIC
 - ☐ mw:SPRO
- Тип оборота** (Type of phrase):
 - ☐ mw:degr.max
 - ☐ mw:dir
 - ☐ mw:dist.max
 - ☐ mw:dur.max
 - ☐ mw:ev
 - ☐ mw:ev.neg
 - ☐ mw:ev.posit
 - ☐ mw:loc.body
- Тип оборота** (Type of phrase):
 - ☐ mw:place
 - ☐ mw:rep
 - ☐ mw:shift
 - ☐ mw:space
 - ☐ mw:speed.max
 - ☐ mw:time
- * При поиске по текстам с неснятой омонимией задание повтора может работать некорректно: день(им.ед) рождения(им.мн или род.ед)**
- Слово перед** (Word before):
 - ☐ любым знаком препинания
 - ☐ точкой
 - ☐ запятой
 - ☐ двоеточием
 - ☐ точкой с запятой
 - ☐ тире
 - ☐ восклицательным знаком
 - ☐ вопросительным знаком
- Слово после** (Word after):
 - ☐ любого знака препинания
 - ☐ точки
 - ☐ запятой
 - ☐ двоеточия
 - ☐ точки с запятой
 - ☐ тире
 - ☐ восклицательного знака
 - ☐ вопросительного знака
- словарное** ☐ / **несловарное** ☐
- слова с заглавной буквы** ☐ / **в начале предложения** ☐ / **в конце предложения** ☐

At the bottom, there are buttons: **OK**, **Очистить** (Clear), and **Отмена** (Cancel).

Как ни странно, *distort* (искажённая форма) — грамматический.

Задание 31. Найдите все прилагательные, написанные с заглавной буквы, но стоящие не в начале предложения. —

Выбор подкорпуса

Вы можете задать подмножество корпуса, по которому в дальнейшем будет вестись поиск. Подробнее о параметрах текста см. в разделе «[Параметры текста](#)».

Омонимия

- ☐ Только тексты со снятой грамматической омонимией **1**
- ☐ Только тексты с неснятой грамматической омонимией

Основные параметры текста ?

Название
 Автор текста **2**
 Пол: ☒ любой ☐ мужской ☐ женский
 Год рождения: с по ☐ Точное вхождение
 Год создания: с **3** по ☐ Точное вхождение

Подкорпусы основного корпуса Иногда нужно исследовать не весь русский язык, а тексты одного периода или конкретного автора (или хочется неслучайным образом ограничить объём работы ввиду ограниченных ресурсов). НКРЯ может выдавать результаты не из всех, а из некоторых текстов в соответствии с условиями на наш выбор. Для этого нужно нажать на **задать подкорпус** на странице поиска. Появится страница выбора.

Когда подкорпус выбран, нужно нажать **Далее >>** внизу страницы, а потом нажать

Сохранить подкорпус и перейти к странице поиска.

- ① Выбрав этот вариант, вы будете искать только по текстам, где эксперты устранили неоднозначные разборы (например, *рыбы в В воде плавают рыбы* получит помету **nom,pl** ‘именительный падеж множественного числа’, а в *Купи немного рыбы* — **gen,sg** ‘родительный падеж единственного числа’). Такие тексты составляют примерно 2 % объёма корпуса.

Задание 32. Ограничившись снятой омонимией, найдите последовательность «не + глагол + точная форма *мышей*». Сколько разборов у последнего слова? Почему? ─

Задание 33. Выберите подкорпус со снятой омонимией и сделайте четыре запроса: слово *лицо* как одушевлённое в единственном числе; одушевлённое во множественном; неодушевлённое в единственном; неодушевлённое во множественном. Какое из них используется чаще в единственном числе, а какое во множественном? ─

- ② Здесь можно написать имя автора, и тогда поиск будет осуществляться только по текстам, написанным этим автором. Берегитесь однофамильцев!

Задание 34. Задайте подкорпус сочинений, написанных автором М. А. Булгаков. Найдите в лексико-грамматическом поиске все употребления слова *Москва* в его сочинениях; затем отдельно найдите все употребления слов *Ленинград* и *Казань*.

Затем задайте подкорпус для автора **Василий Гроссман** и найдите те же три слова. Затем задайте подкорпус для автора **А. И. Солженицын** и сделайте то же самое. Кто из авторов чаще упоминает Казань, чем Ленинград? ─

- ③ Здесь можно написать период, тексты которого нас интересуют, например с **1851** по **1900**; **Точное вхождение** означает, что в диапазон не попадёт текст, написанный, например, в период с 1897 по 1903 год.

Задание 35. Задайте подкорпус текстов **со снятой омонимией**, написанных с 1701 по 1800 г. (с точным вхождением). Найдите в них все употребления прилагательных в форме женского рода, творительного падежа, которая оканчивается на *...ю* (“*й” в поле **Слово**). Затем отдельно найдите все употребления прилагательных в той же форме, оканчивающихся на *...ю*.

Повторите эти два запроса для текстов, написанных с 1801 по 1900 г., а потом для текстов, написанных с 1901 по 2000 г. ─

Кроме того, можно выбирать тексты по жанру, типу текста (роман, повесть, пьеса и т. д.), для нехудожественных текстов (non-fiction) — по тематике и т. д.

Другие корпуса НКРЯ Кроме основного, в НКРЯ есть другие корпуса (перечислены слева на странице поиска). Например, в поэтическом корпусе можно искать по стихотворным текстам, а в устном — по текстам, записанным с голоса.

Поэтический корпус: важное свойство — знаки грависа (`) на сильных долях стиха. В реальности там может быть пиррихий (а на слабой доле — спондей), и часты слова с несколькими сильными иктами, но всё-таки это кое-что даёт. Можно ограничить поиск слова «зоной рифмовки». Поиск по метру, длине строки в иктах, рифмовке...

Задание 36. Задав подкорпус с парной или тройной рифмой, получите выдачу зарифмованных глагольных форм 3 лица ед. ч. типа *Ветер по морю гуляет / И кораблик подгоняет*. Определите, как меняется отношение поэтов к такой рифме со временем: задайте по такому подкорпусу для XVIII, XIX, XX и XXI вв.; подсчитайте для каждого из периодов отношение числа **документов** с такими рифмами к числу документов в данном подкорпусе и постройте график изменения этого отношения со временем. Годятся только такие рифмы, где зарифмованы две или три такие формы **друг с другом!** Нужно найти такие случаи с минимальным числом ошибочных примеров в выдаче. ─

Синтаксический корпус позволяет искать слова и сочетания слов с учётом синтаксических отношений, в которые они вступают.

Основные подходы к описанию синтаксиса:

через зависимости: отношение «главного» слова и «навешенных» на него

через (непосредственно) составляющие: отношение слова, определяющего свойства синтаксического единства, и прочих элементов этого единства. Единство наследует свойства одного из своих элементов.

(1) Я [опять [вспомнил [[произошедшие вчера] событиях]]].

Большая теоретическая и практическая проблема для русского языка — разрывные составляющие (как проявление «свободного порядка слов»)



При анализе в зависимостях в узлах дерева словоформы, а при анализе в составляющих — составляющие разного размера, т. е. дерево зависимостей отражает что-то вроде передачи права находиться в предложении от слова к слову, а дерево составляющих — отношения «часть—целое» между фрагментами предложения.

Школьная грамматика ближе к зависимостям, если не считать целикомого маркирования причастных и деепричастных оборотов. Фактически школьная система, включая [двойные разборы](#), реализована в парсере ХАНКО.

Задание 37. Попробуйте проанализировать двумя способами предложение (2).

(2) Олицетворением их был, в частности, С. Трапезников, долгое время возглавлявший Отдел науки и учебных заведений ЦК КПСС.

На каком подходе основан его [анализ](#) в ЭТАП-3 / «СинТагРус»? ─

В синтаксическом корпусе НКРЯ не работает поиск латиницей! Но можно выбирать значения признаков мышкой из списка. Кроме того, в нём иначе понимается расстояние: это расстояние от данного слова до того, от которого оно зависит в синтаксисе. Чтобы искать несколько слов, можно воспользоваться кнопкой , чтобы создать поле для ещё одного слова, подчинённого вместе с данным тому, от которого данное зависит, или же кнопкой , чтобы создать поле для слова, подчинённого данному.

Задание 38. Сравните результаты запросов *V, ном* и *V, им*. ─

Задание 39. Найдите все цепочки из четырёх зависящих друг от друга существительных ср. р. Все ли результаты подходят под определение зависимости? При чём тут примеры про *грибы* из [руководства](#) по использованию синтаксического корпуса? ─

Задание 40. Найдите все сочетания трёх слов с вершиной-существительным и двумя зависимыми от него существительными (в родительном и творительном падежах) типа *бомбардировка урана нейтронами*, не указывая расстояния от родителя. Сравните выдачу с выдачами в основном подкорпусе при различных указанных расстояниях. ─

Синтаксические отношения можно аннотировать по семантике связи. В НКРЯ, однако, подход более формально-ориентированный; во внимание принимаются обязательность заполнения валентности (ср. комплетивные отношения) и частеречная принадлежность связываемых слов. Более семантический подход реализован в [каталогах фреймов](#).

Параллельный корпус — это корпус, где собраны переводы текстов на разные языки. При этом языком оригинала или перевода должен быть русский.

Можно искать и по русским словам, и по иностранным. Корпус для каждого найденного употребления покажет фрагмент текста, соответствующий ему на другом языке. Теперь можно [искать](#) сразу по двум языкам, так что в конкордансе искомые слова будут подсвечены на обоих языках.

Задание 41. Выберите английский параллельный корпус и найдите в нём все употребления сочетания *дурак дураком* (можно воспользоваться поиском точных форм). Какими выражениями переводится это выражение на английский язык? ─

Задание 42. Выберите английский параллельный корпус и найдите в лексико-грамматическом поиске сочетание «*либо* + соседнее существительное + *либо* + существительное (соседнее или через одно слово)». ─

Задание 43. Задайте **подкорпус** многоязычного корпуса, состоящий из текстов, переведённых с любого языка на латинский (см. рисунок). Установите, какое выражение соответствует в английском языке слову *козявка* в русском тексте из этого корпуса. ─

главная

Параллельный корпус: выбор подкорпуса

Вы можете задать подмножество корпуса, по которому в дальнейшем будет вестись поиск. Подробнее о параметрах текста см. в разделе «[Параметры текста](#)».

Корпус: ☐ двуязычный ☒ многоязычный

Название

Автор текста [выбрать](#)

Год рождения: с по ☐ Точное вхождение

Год создания: с по ☐ Точное вхождение

Переводчик

Язык оригинала:

Язык перевода:

☐ Показать обе

Для выполнения следующего задания нужно нажать ← Вернуться к старой версии сверху на странице поиска, а затем перейти в параллельный корпус.

Задание 44. Найдите в лексико-грамматическом поиске многоязычного параллельного корпуса все употребления слова *lass*, указав, что оно интересует нас в текстах на английском языке (**Дополнительные признаки и языки**). Сравните результат с результатом запроса, где вместо этого потребован немецкий язык.

Найдите в том же режиме поиска все употребления слова *брат*; сравните результат с результатом запроса того же слова, при котором из языков **исключён** русский язык (**Дополнительные признаки и языки**, выбрать русский, затем поставить «-»: -ru). ─

5 Интернет-корпусы. CQL

Aranea Благодаря Интернету возможно автоматическое составление корпусов (Web as Corpus). Они создаются путём чистки текстов, собранных в Сети алгоритмом-«пауком» (crawler). Существует несколько таких проектов для русского языка, в частности Aranea.

Задание 45. Познакомьтесь с интерфейсом NoSketch Engine для корпуса [Araneum Russicum Minus](#). Сравните результаты выдачи для одного и того же запроса в полях Lemma и Word form. На примере символа *н* изучите поведение поля Character. На классическом примере *дали* познакомьтесь с механизмом уточнения запроса в поле Word form указанием части речи.

Поле CQL предназначено для запросов, написанных на Corpus Query Language.

Имеется [перечень](#) грамматических меток для русского языка. Результаты поиска можно скачать (Save слева; до 100000 примеров); XML можно затем открывать с помощью Excel.

[word="пузо"] точная форма *пузо*

[lemma="пузо"] все словоформы лексемы *пузо*

[tag="Npfsgy"] все вхождения одушевлённых собственных существительных женского рода в формах ед. ч. род. п.

[word="Хо.ский"] любая буква на месте пропуска

[word="Хо[дф]ский"] любая буква из данных на месте пропуска

[word="Хо[мд]ский"] любая буква, кроме данных, на месте пропуска

[word="воо*т" & word!="вот"] любое количество *о* (не 0, это мы исключили после &)

[lemma=".*густ" & lemma=".*август" & lemma!=".*Август"] все словоформы слов, в начальной форме заканчивающихся на ...*густ*, кроме форм слов *август*, *Август* и сочетаний без пробела с ними на конце

[tag="N...y"] все вхождения одушевлённых существительных

[word="времен"] []{1,3} [word="покоренья"] от одного до трёх слов между крайними элементами запроса (знаки препинания — тоже «слова»: \., \, и т. д.)

[word="своей"|word="своею"] [word="собственной"]{0,1} [word="рукой"] слово посередине, которого может не быть в выдаче, и дизъюнкция признаков в первом

1: [lemma="птица"] 2: [] & 1.tag=2.tag по два слова с полным совпадением грамматических признаков, причём первое слово — форма лексемы *птица*

Задание 46. Найдите в Araneum Russicum Minus все случаи употребления конструкции типа *шутки шутками*, *а...* (падежи и повтор лексемы важны, конкретное существительное может быть любым).

Задание 47. Найдите в [списке](#) корпусов Aranea корпус Minus изучаемого языка и найдите там предложения, соответствующие русским предложениями структуры «существительное-подлежащее + переходный глагол-сказуемое в личной форме + существительное-дополнение».

Корпусы Университета Лидса British National Corpus (BNC) — один из самых известных в мире и самых ранних общедоступных корпусов. Содержит 100 млн словоупотреблений из текстов примерно тридцатилетней давности. Он доступен на нескольких сайтах, например [здесь](#) с регистрацией (и подробной [инструкцией](#)) или [здесь](#) с упрощённым языком запросов (и тоже с регистрацией). Мы рассмотрим его [версию](#) на сайте Университета Лидса.

reopens или [word="reopens"] точная форма *reopens*
my dearest сочетание форм *my dearest*

[lemma="reopen"] все формы *reopen*
[lemma=".o[s1]e"] *role, lose, posed...*
[lemma="cooperator|nasal"|word="surfed"] дважды используется ‘или’
[word="my"] [word!="\w"]{2,3} [word="family"] от двух до трёх слов в середине
[lemma="bring"&pos="VVD"] [word="me"] [pos="IN"&word!="to"] *brought me into, ...*

Задание 48. Найдите в BNC все последовательности, где за сочетанием *our sinful* непосредственно следует **любая** форма существительного. —

Задание 49. Найдите в BNC все последовательности, где за любой формой слова *call* следует *the*, а затем любая форма любого из слов *police, sheriff* или *FBI*. —

Задание 50. Найдите в BNC все сочетания из двух слов подряд, где первое — слово из шести букв, оканчивающееся на *-ely*, а второе — любое, оканчивающееся на *-ish*. —

Задание 51. Назовите единственное сочетание, которое будет найдено в BNC по запросу «*skeptical* или *sceptical* + *and* + одно любое слово + любая форма слова *attitude*». —

Задание 52. Найдите в BNC все сочетания, в которых первое слово **не** *Sherlock* (‘не’ не будет работать без `pos="..."&`), а второе — *Holmes*. Повторите поиск, дополнительно потребовав, чтобы первое слово было именем собственным в ед. ч. (singular proper). —

В разделе **Collocation** можно выяснить, какие слова чаще всего сочетаются с данной последовательностью слева или справа.

Задание 53. Составьте список прилагательных, сочетающихся слева с *and tired*. —

Создание корпуса На сайте университета Лидса можно создать собственный корпус (в т. ч. для русского языка!) с морфологической разметкой. Для этого на начальной странице нужно выбрать **Build or Search Your Own Corpora**, а затем зарегистрироваться или войти. Затем можно загрузить до пяти файлов *.txt*, *.doc* или *.zip*. На вкладке **Build Corpus** можно выбрать файлы для включения в корпус, название корпуса и язык (важно для разметки), на последней вкладке — добавить корпус в список корпусов выбранного языка.

К полученному корпусу применим поиск на языке запросов, включая Search Builder с возможностью задать грамматические признаки мышкой (PoS Editor).

Задание 54. Создайте корпус *myown* на русском языке, используя три больших файла с вашими собственными текстами.

Найдите в нём все существительные женского рода в дательном падеже. —

6 Обработка данных корпуса

Задание 55 (повторение приёмов работы с Excel). Создайте таблицу: в первом столбце фамилия студента (используйте готовый список из ≥ 5 вымышленных фамилий, например персонажей Стругацких, аниме, хоббитов и т. п.), в следующих трёх оценки за три контрольных работы по пятибалльной системе, в пятом столбце сумма оценок данного студента (функция **СУММ**), в шестом **доля**, которую суммарная оценка данного студента составляет от **максимальной суммы в данной таблице**. Написав формулу для первого студента в списке (понадобится функция **МАКС**), растяните содержимое этой ячейки в клетки с долями для других студентов. Закрепите координаты ячеек, на которые ссылается формула в ячейке с долей, так, чтобы при растягивании максимум продолжал рассчитываться для той же совокупности студентов.

Отсортируйте таблицу по фамилиям студентов; по суммарной оценке. —

Корпус, например НКРЯ, умеет искать по самым разным признакам, но всё-таки это не все признаки, по которым хотим искать мы. Кроме того, в корпусе бывают и ошибки. Поэтому результаты поиска иногда нужно скачать на свой компьютер и обрабатывать в виде таблицы, например в Microsoft Excel или Google Sheets.

Например, мы хотим изучить, от чего зависит порядок слов в русском языке в предложениях с глаголом *дать*: можно сказать *Маша дала карандаш Пете*, а можно — *Маша дала Пете карандаш*. Но как говорят чаще? Возможно, это зависит от длины слова: если слово в винительном падеже короткое, а слово в дательном длинное, слово в дательном будет в конце, а если наоборот — вначале?

Корпус даст нам возможность проверить это предположение, но найти слова определённой длины в нём нельзя. Поэтому придётся найти все примеры с *дать* и двумя существительными, а потом вручную разделить их на группы в зависимости от того, какую длину имеют в них слова.

Задание 56. Задайте в НКРЯ подкорпус текстов, созданных начиная с 2001 года. Найдите там все сочетания, в которых сначала идёт глагол *дать* (в любой форме изъявительного наклонения), затем без знака препинания идёт существительное в винительном падеже, а затем без знака препинания существительное в дательном падеже, за которым следует знак препинания. (Знак препинания в конце нам нужен, чтобы у последнего слова не было своих зависимых слов.)

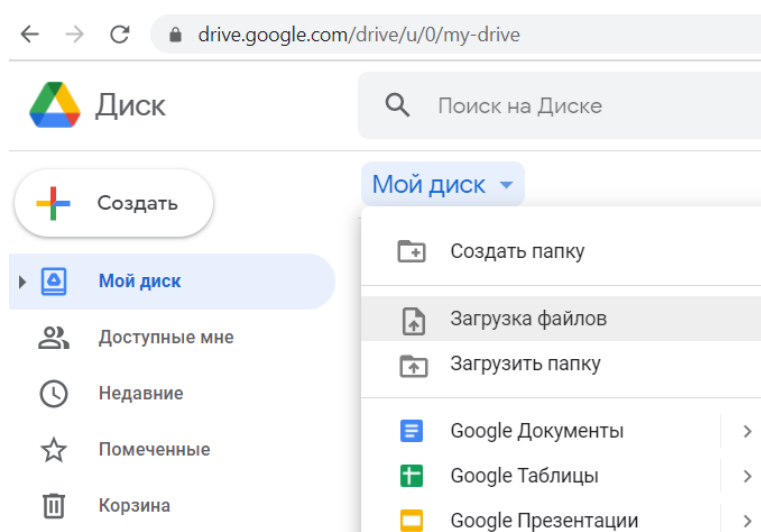
Внизу на странице результатов есть команда **Скачать несколько первых результатов выдачи в формате Excel**. Если результатов немного (несколько сотен), скачиваются все.

Задание 57. Скачайте только что найденные результаты в формате Excel. Назовите этот файл **ACCDAT.xlsx**.

Найдите в том же подкорпусе сочетания, в котором порядок винительного и дательного падежа противоположный. Скачайте файл Excel и назовите его **DATACC.xlsx**.

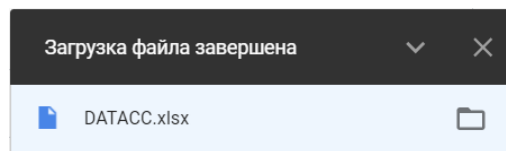
Открыть полученные файлы можно в разных программах, например в Microsoft Excel или [веб-версии Excel](#)¹, в бесплатной программе [LibreOffice Calc](#) или в Google Sheets.

Рассмотрим последний вариант. После регистрации на сайте <http://drive.google.com> вы сможете загрузить файл с компьютера:

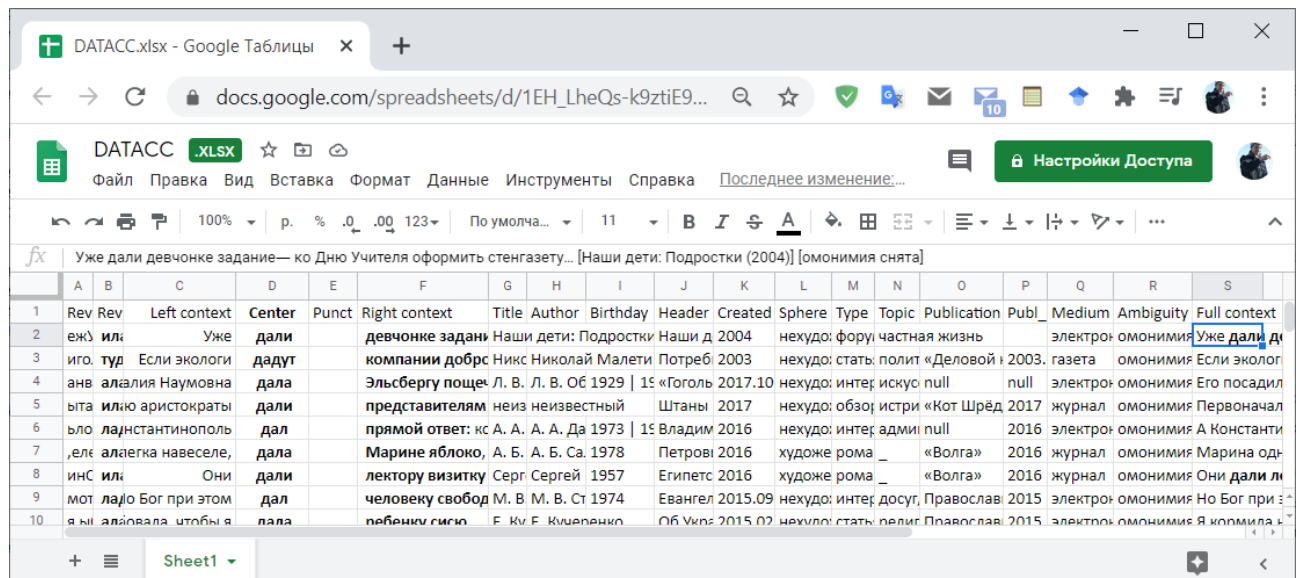


Когда файл будет загружен, появится окно, куда можно нажать, чтобы открыть файл:

¹Доступна после регистрации на сайте <http://office.com> в разделе Продукты > Для учащихся и преподавателей; используйте адрес электронной почты типа `stXXXXXX@students.spbu.ru`.

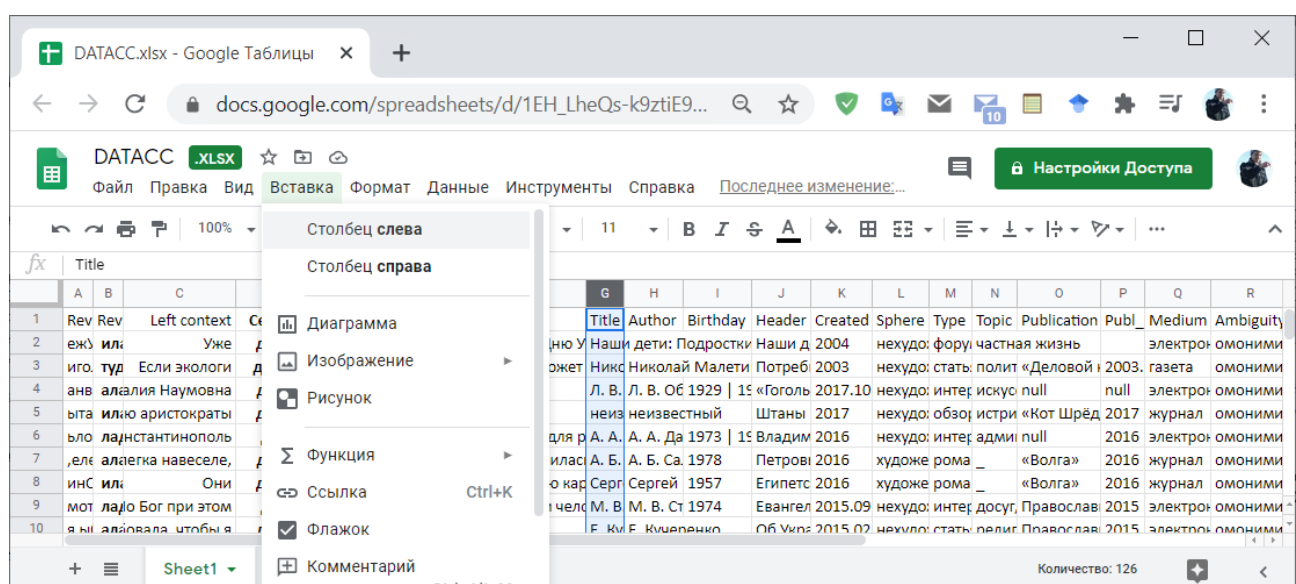


Откроется просмотр, но сверху будет кнопка Открыть в приложении Google Таблицы. Если нажать на неё, таблица откроется для редактирования.



В столбце Center находится первое слово, которое мы искали (форма глагола *дать*), столбец Right context начинается с остальных слов, которые мы искали. Примеры целиком находятся в последнем столбце Full context.

Теперь можно начать группировать примеры. Для этого заведём три новых столбца (columns): в первом (DAT) будем писать длину слова в дательном падеже (количество слогов в нём), во втором (ACC) — количество слогов в слове в винительном падеже. В третьем столбце будем считать разность (Difference): если она больше 0, то длиннее слово в дательном падеже, если меньше 0 — длиннее слово в винительном падеже.



Создав столбцы, начнём вписывать в DAT и ACC длину слов в слогах, а в Difference напишем **формулу**, которая будет подсчитывать разность. Формула начинается со знака =, за которым идёт вычитание («-») числа в столбце ACC из числа в DAT.

DATAACC.xlsx - Google Таблицы

docs.google.com/spreadsheets/d/1EH_LheQs-k9ztiE9...

DATAACC .xlsx

Файл Правка Вид Вставка Формат Данные Инструменты Справка Последнее изменение...

100% 11 B I S A

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Rev	Rev	Left context	Center	Punct	Right context	DAT	ACC	Difference	Title	Author	Birthday	Header	Created	Sphere	Type	Topic	Publication	Put
2	еж\	ил\	Уже	дали		девчонке задание — ко Дню У	3	4	-1	Наши дети: Подростки Наши д	2004				нехудо\	фору\	частная жизнь		
3	иг\	тур	Если экологи	дадут		компании добро, она сможет	4	2	2	Никс Николай Малети	Потреб	2003			нехудо\	статы\	полит «Деловой	200	
4	анв	алалия	Наумовна	дала		Эльсбергу пощечину.	3	4	=G4-H4	Л. В. Л. В. Об	1929	15	«Гоголь	2017.10	нехудо\	интер\	искус\	null	nul
5	ыта	илию	аристократы	дали		представителям черни.				неиз	неизвестный		Штаны	2017	нехудо\	обзор\	истри «Кот Шрёд	201	
6	бло	ла\	нстантинополь	дал		прямой ответ: комиссия для решения этого				А. А. А. А. Да	1973	15	Владим	2016	нехудо\	интер\	адми\	null	201
7	,ел\	алаетка	навеселе,	дала		Марине яблоко, выговорились, даже не спросив,				А. Б. А. Б. Са.	1978		Петров	2016	худож\	рома _	«Волга»	201	
8	инс	или	Они	дали		лектору визитку — пустую карточку с телефоном				Серг\	Сергей	1957	Египетс	2016	худож\	рома _	«Волга»	201	
9	мот	ла\	Бо\	при этом	дал	человеку свободу, и если человек отказывается				М. В. М. В. Ст	1974		Евангел	2015.09	нехудо\	интер\	досуг, Православ	201	
10	я н\	ала\	вала,	чтобы я	дала	ребенку гиско.				Е. К\	Е. К\	К\	К\	2015.02	нехудо\	статы\	религ\	Православ	201

Sheet1

Анализ данных

Чтобы одна формула для вычитания распространилась на весь столбец, нужно поставить курсор в правый нижний угол (он превратится в +) и «потянуть» ячейку вниз:

G	H	I	J
DAT	ACC	Difference	Title
3	4	-1	Наши
4	2		Никс
3	4		Л. В.
			неиз
			решения этого
			ь, даже не спросив,
			А. А.
			А. Б.

Некоторые примеры (как *представителям черни* в строке 5 или *прямой ответ* в строке 6 на рисунках выше) найдены неправильно: в них существительные стоят в неправильном падеже или не соответствует запросу что-то ещё. Такие строчки можно удалить:

DATAACC.xlsx - Google Таблицы

docs.google.com/spreadsheets/d/1EH_LheQs-k9ztiE9...

DATAACC .xlsx

Файл Правка Вид Вставка Формат Данные Инструменты Справка Последнее изменение...

Вставить Ctrl+V

По умолчанию 11 B I S A

Специальная вставка

Переместить строку вверх

Переместить строку вниз

Найти и заменить Ctrl+N

Удалить значения

Удалить строку 5

Удалить примечания

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Rev	Rev	Left context	Center	Punct	Right context	DAT	ACC	Difference	Title	Author	Birthday	Header	Created	Sphere	Type	Topic	Publication	Put
2	еж\	ил\	Уже	дали		девчонке задание — ко Дню У	3	4	-1	Наши дети: Подростки Наши д	2004				нехудо\	фору\	частная жизнь		
3	иг\	тур	Если экологи	дадут		компании добро, она сможет	4	2	2	Никс Николай Малети	Потреб	2003			нехудо\	статы\	полит «Деловой	200	
4	анв	алалия	Наумовна	дала		Эльсбергу пощечину.	3	4	-1	Л. В. Л. В. Об	1929	15	«Гоголь	2017.10	нехудо\	интер\	искус\	null	nul
5	ыта	илию	аристократы	дали		представителям черни.				неиз	неизвестный		Штаны	2017	нехудо\	обзор\	истри «Кот Шрёд	201	
6	бло	ла\	нстантинополь	дал		прямой ответ: комиссия для решения этого				А. А. А. А. Да	1973	15	Владим	2016	нехудо\	интер\	адми\	null	201
7	,ел\	алаетка	навеселе,	дала		Марине яблоко, выговорились, даже не спросив,				А. Б. А. Б. Са.	1978		Петров	2016	худож\	рома _	«Волга»	201	
8	инс	или	Они	дали		лектору визитку — пустую карточку с телефоном				Серг\	Сергей	1957	Египетс	2016	худож\	рома _	«Волга»	201	
9	мот	ла\	Бо\	при этом	дал	человеку свободу, и если человек отказывается				М. В. М. В. Ст	1974		Евангел	2015.09	нехудо\	интер\	досуг, Православ	201	
10	я н\	ала\	вала,	чтобы я	дала	ребенку гиско.				Е. К\	Е. К\	К\	К\	2015.02	нехудо\	статы\	религ\	Православ	201

Количество: 18

Анализ данных

Так можно подсчитать для каждого примера, какое из существительных в нём длиннее. Потом мы можем посмотреть на все числа в столбце **Difference** и посчитать, сколько из них меньше 0, сколько 0, а сколько больше 0.

То же самое можно сделать для файла **ACCDAT.xlsx** (не забудем там тоже вычитать длину слова в **винительном** падеже из длины слова в **дательном** падеже, хотя они стоят в обратном порядке; иначе говоря, формула будет «=H4-G4»).

Когда подсчёты закончены, можно обобщить результаты. У нас получается, что разница в длине между словом в дательном падеже и словом в винительном падеже больше, если дательный стоит после винительного: «что длиннее, то и в конце», возможно, работает, но надёжность этого вывода нужно бы проверить методами математической статистики.

DAT – ACC = ...	DATAcc.xlsx	ACCDAТ.xlsx
–4	1	
–3	5	
–2	6	13
–1	28	21
0	22	32
1	23	18
2	14	33
3	8	12
4	3	5
5		1
<hr/>		
< 0	40	34
= 0	22	32
> 0	48	69
<hr/>		
среднее	0,25	0,73

Таблица 1: Длина существительных в винительном и дательном падежах в примерах из заданий 56 и 57 (средний столбец — таблица на рисунках выше)

Рассмотрим другой пример. Известно, что теоретически род аббревиатуры определяется по стержневому слову в сочетании, от которого она образуется; например, *МГУ* м. р., *ООН* ж. р., *МЦК* ср. р. (Московское центральное кольцо). В речи мы иногда нарушаем это требование. Исследуем реальную практику согласования прилагательных с аббревиатурами, обозначающими организации и теоретически относящимися к м. р.

Задание 58. Запросите все вхождения сочетания «прилагательное в ед. ч., в И. п. или В. п. (но не в омонимичном ему Р. п.) + несклоняемое (грамматический признак 0) название организации (семантический признак **t:org**) м. р. (но не ж. р. и не ср. р.)». Это сочетания типа *местный КГБ* или *сегодняшнее ООН*. Создайте выгрузку результатов. Отсортируйте результаты в Excel по правому контексту (т. е. по искомому существительному). Пометьте строки, где найденное слово не является аббревиатурой (скажем, несклоняемое имя) или не м. р. (как *ПТУ*). Подумайте, почему удобнее делать сортировку до отсева irrelevantных строк. Скопируйте лист и удалите эти строки из **копии**. Пометьте строки, где синтаксическое отношение между **Center** и **Right context** — это не искомое нами согласование (скажем, *рядовой КМП* (Корпуса морской пехоты США), где связь — не согласование, а управление и где она направлена не от аббревиатуры, а к ней). Скопируйте лист и удалите эти строки из копии листа.

Не удаляйте примеры с «неправильным» родом прилагательного типа *местное КГБ!* Это как раз то, что мы изучаем!

Вставьте пустой столбец перед **Title**, назовите его **A gender** и начните размечать выдачу: вписывайте в **A gender** помету **m**, если слово в **Center** имеет форму м. р., **n**, если ср. р., и **f**, если ж. р.

Отсортировав все неудалённые строки по столбцу **A gender**, подсчитайте число случаев согласования по м. р., по ж. р. и по ср. р. Постройте круговую диаграмму, отражающую доли трёх вариантов согласования от общего числа случаев, оставшихся в таблице.

└

О языковом разнообразии Когда появился Интернет, стало возможно собрать в одном месте информацию о языках мира. Сайт «[Glottolog](#)» даёт общие сведения о языках Земли. Как и WALS (о котором мы будем говорить на следующей неделе), он относится к группе ресурсов, создаваемых институтами Общества Макса Планка в Германии. В разделе «Languages» можно искать языки по названию, языковой семье (**Top-level family**; только для языков, у которых известны родственники), части света (**Macro-area**).

Если щёлкнуть на название языка, можно просмотреть сведения о его генетической классификации (1), сохранности (2), географическом расположении (3) и список литературы об этом языке (4). Например, для айнского языка, на котором говорили в Японии:

25

Задание 59. Пользуясь поиском по языкам, определите, (а) есть ли родственники у языка *пираха* (Pirahã) и (б) считается ли английский язык индоевропейским языком Австралии (и какие вообще языки считаются таковыми); для этого выберите часть света **Australia** и семью **Indo-European**. ↵

В разделе «Families» можно искать по языковым группам и семьям. В столбце **Level** можно выбрать, будут ли найдены группы (**Subfamily**), более крупные объединения — семьи (**Top-level family**), языки без известных родственников (**Isolate**) или семьи вместе с изолятами (**Top-level unit**).

Задание 60. Определите, какие искусственные языки (**artificial languages** — условная «семья») им учитываются. Где на карте локализован язык *эсперанто*? Найдите в списке литературы самую новую книгу об этом языке, известную сайту «Glottolog». ↵

Другой ресурс, где можно получить информацию о языках мира, — это «[Ethnologue: Languages of the World](#)», но там почти ничего нельзя сделать без регистрации (в отличие от «Glottolog», это коммерческий продукт). В некоторых странах, например в России, регистрация бесплатная, и после неё можно посмотреть данные о языках (**Languages**) и о языковой ситуации в странах (**Countries**). Так выглядит часть страницы русского языка (**Classification** — генетическая классификация, т. е. языковое родство, от семьи к подгруппе; **Typology** — основные черты фонологии, морфологии, синтаксиса):

Russian



LANGUAGE

MAP

A language of Russian Federation

ISO 639-3 [rus](#)

Autonym русский язык (russkij jazyk)

User Population 138,000,000 in Russian Federation (Arefyev 2012), all users. L1 users: 119,000,000 (Arefyev 2012). Total users in all countries: 257,962,060 (as L1: 153,580,550; as L2: 104,326,510).

Language Status 1 (National). Statutory national language (1993, Constitution, Article 68(1)). Provincially recognized language in Dagestan Autonomous Republic (1994, Constitution of Dagestan Autonomous Republic, Article 10).

Classification [Indo-European](#), [Balto-Slavic](#), [Slavic](#), [East](#)

Dialects North Russian, South Russian.

Typology SVO; prepositions; genitives after noun heads; adjectives, numerals before noun heads; question word initial; 1 prefix on a word; recursive addition of suffixes allowed; gender (masculine/feminine/neuter); no articles; case-marking (6 cases); verb affixes mark person, number; passives; tense and aspect; comparatives; 32 consonants, 5 vowels, 4 diphthongs; non-tonal; free stress.

А в профиле страны можно посмотреть данные о всех языках, на которых в ней говорят (причём по группам, соответствующим разным статусам — есть ли на этих языках литература, обучение и т. д.); например, так выглядит начало списка для России:

Russian Federation

 Print

COUNTRY

LANGUAGES

STATUS

MAPS 

Expand All

Collapse All

1 (National)

Hide Details 

Russian

[rus] 1 (National). Statutory national language (1993, Constitution, Article 68(1)). Provincially recognized language in Dagestan Autonomous Republic (1994, Constitution of Dagestan Autonomous Republic, Article 10). 138,000,000 in Russian Federation (Arefyev 2012), all users. L1 users: 119,000,000 (Arefyev 2012). Total users in all countries: 257,962,060 (as L1: 153,580,550; as L2: 104,326,510).

2 (Provincial)

Hide Details 

Adyghe



[ady] 2 (Provincial). Statutory provincial language in Adyghea Republic (1995, Constitution, Adyghea Republic, Article 2). 117,500 in Russian Federation (2010 census). No monolinguals (Ministry of Education, Adyghea Republic). Ethnic population: 129,000 (2010 census). Total users in all countries: 568,500.

Altai, Southern

[alt] 2 (Provincial). 57,400 (2010 census). Ethnic population: 74,200 (2010 census). Includes Northern Altai [atv].

О языковых явлениях Базы данных в Интернете могут быть посвящены не только языкам мира, но и языковым явлениям, таким как порядок слов, наличие или отсутствие падежей, одно или разные наименования для руки и для кисти руки и т. д. В них можно увидеть, какие признаки и какие их комбинации встречаются чаще (в большом количестве языков), а какие не встречаются вообще.

Фонологические особенности языков мира собраны в базе данных «Phoible». Там в разделе «Inventories» можно посмотреть, какие фонемы есть в интересующем нас языке. Бывает, что для одного языка приводится несколько описаний по разным источникам:

['fɔɪ.bɪl]							
Home Contributors Inventories Languages Segments Sources							
Inventories							
Showing 1 to 2 of 2 entries (filtered from 3,020 total entries)				← Previous 1 Next →		 	
Inventory	Language	# segments	# vowels	# consonants	# tones	Contributor	Cite
Hawai	Search		Search	Search	Search	--any--	
Hawaiian (SPA 43)	Hawaiian	19	10	9	0	Stanford Phonology Archive	cite
HAWAIIAN (UPSID 352)	Hawaiian	13	5	8		UCLA Phonological Segment Inventory Database	cite

На странице языка даётся таблица его фонем; Representation — доля всех языков на сайте, в которых есть такая же фонема. Вкладка IPA Chart располагает те же фонемы в стандартных таблицах Международного фонетического алфавита (его можно слушать).

Задание 61. Найдите в базе «Phoible» описания фонологии русского языка. Найдите среди них то, в котором /ы/ ([i]) считается отдельной фонемой. Назовите самые распро-

странённые в языках мира (согласно столбцу **Representation**) и самые редкие фонемы в этом описании. ─

В разделе «Segment» можно искать отдельные фонемы, точнее знаки, которые используются для их записи. Если вам в тексте о каком-нибудь языке попался незнакомый знак, можно попытаться скопировать его и вставить в поле **Name**. В столбце **Representation** можно искать определённое число, например 43, но можно искать <43 или >43. Поле **Segment class** можно выбрать, что мы будем искать: что угодно, только гласные, только согласные или только тоновые фонемы.

Грамматические (но также фонологические и лексические) особенности языков собраны в масштабной базе «[World Atlas of Language Structures](#)» (WALS). Она следует современным стандартам типологического описания языков, при котором различия между языками не сглаживаются, но и не усиливаются за счёт терминов, с помощью которых мы описываем эти языки.

WALS состоит из [карт](#), показывающих, как языки мира распределяются по возможным значениям того или иного параметра (например, по порядку слов в предложении или по количеству родов), и [пояснительных статей](#), написанных специалистами по соответствующим вопросам. Таким образом, это не только атлас, но и энциклопедия.

Задание 62. Найдите в WALS признак (feature) 138A «Tea». Какие языки Европы не попадают ни в одну из двух основных групп? ─

Задание 63. В разделе «Features» сайта WALS найдите признак 86A «Order of Genitive and Noun». Что это значит? К какому типу относится русский язык? К каким типам относятся арабский, японский языки? Можно ли сказать, что один из типов включает большинство языков? (Ср. эту ситуацию с ситуацией для признака 81A «Order of Subject, Object and Verb».) ─

Задание 64. В разделе «Languages» WALS найдите китайско-русский пиджин. На скольких картах WALS появляется этот идиом? По каким из этих признаков он совпадает с русским? А с китайским? ─

Задание 65. Найдите в WALS признак 136A «M-T Pronouns». Добавьте к карте признак 13A «Tone» и нажмите **Submit**. Назовите единственный язык, где есть «парадигматические» местоимения М-Т и вдобавок сложная система тонов. ─

Универсалии (общие черты всех языков) и раритеты собраны в [Архиве универсалий](#) (несколько устаревшее [введение](#)).

Задание 66. Найдите универсалии и раритеты из области синтаксиса, в чьей исходной (original) формулировке упоминается русский язык (Russian). Приведите пример явления, о котором говорится в единственном найденном раритете, и явления, о котором идёт речь в универсалии с самым большим номером из найденных. ─

Заемствования из одного языка в другой собраны в базе «[World Loanword Database](#)»; её создатели изучили около 40 языков и выяснили, слова с каким значением и из каких языков в них заимствованы. В [списке языков](#) базы те языки, заимствования в которые изучались, раскрашены в разные цвета, а на карте представлены красными значками.

Задание 67. Найдите в «World Loanword Database» японский язык как реципиент (заимствующий язык). Для этого в разделе «Languages» нажмите на цветную клетку с японским языком в таблице или найдите японский в разделе «Vocabularies»:

← Previous 1 Next →

ID	Vocabulary	Author	Number of words	Percentage of loanwords	Cite
<input type="text" value="Search"/>	<input type="text" value="ese"/>		<input type="text" value="Search"/>	<input type="text" value="Search"/>	
21	Japanese	Christopher K. Schmidt	1975	36%	cite
22	Mandarin Chinese	Thekla Wiebusch	2049	2%	cite
24	Vietnamese	Mark Alves	1477	27%	cite

Поиск литературы Если вы хотите о чём-нибудь написать тезисы доклада, статью или дипломную работу, нужно узнать, что об этом написали другие.

Академия Google — важный ресурс для поиска научных публикаций. В строке поиска можно написать термин или имя автора, который вас интересует. У найденных книг и статей справа может быть ссылка PDF: это значит, что можно прочитать текст целиком, если нажать на эту ссылку.

Эмбик и Нойер

Академия

Результатов: примерно 2 750 (0,04 сек.)

Мovement operations after syntax

[PDF] psu.edu

D Embick, R Noyer - Linguistic inquiry, 2001 - MIT Press

We develop a theory of movement operations that occur after the syntactic derivation, in the PF component, within the framework of Distributed Morphology. The theory is an extension of what was called Morphological Merger in Marantz 1984 and subsequent work. A primary ...

☆ 99

Цитируется: 1402

Похожие статьи

Все версии статьи (18)

Если нажать на кнопку «”» под названием источника, появится окно с библиографической ссылкой для этого источника. Ссылку типа «ГОСТ» можно вставить в дипломную работу, а другие типы ссылок подойдут для публикаций на английском языке.

Если нажать на ссылку **Цитируется**, получится список статей и книг, у которых в списке литературы есть тот источник, который мы искали сначала.

Если у вас есть учётная запись (account) Google, можно нажать на ☆ (превратится в ★), чтобы добавить источник в «мою библиотеку», чтобы не забыть. На главной странице Академии есть ссылка **Моя библиотека**, где можно посмотреть, что вы добавили раньше.

✕
Цитировать

ГОСТ Embick D., Noyer R. Movement operations after syntax //Linguistic inquiry. – 2001. – Т. 32. – №. 4. – С. 555-595.

MLA Embick, David, and Rolf Noyer. "Movement operations after syntax." *Linguistic inquiry* 32.4 (2001): 555-595.

APA Embick, D., & Noyer, R. (2001). Movement operations after syntax. *Linguistic inquiry*, 32(4), 555-595.

[BibTeX](#) [EndNote](#) [RefMan](#) [RefWorks](#)

Задание 68. Пользуясь Академией Google, напишите библиографическую ссылку в формате ГОСТ на статью «К типологии относительного предложения». ─

Задание 69. Найдите единственный источник на русском языке, цитирующий статью “Choice-functional series of indefinite pronouns and Hamblin semantics” и написанный не ранее 2016 года. Напишите ссылку на него в формате ГОСТ. ─

Задание 70. Используя команду поиска **author:N Chomsky**, выясните, о чём пишет человек с таким именем. Затем с помощью той же команды найдите работы, «автором» которых является **Незнайка**. Как получаются такие ошибки? ─

Часто авторы публикуют свои работы в Интернете раньше, чем они будут опубликованы в журнале. Такие предварительные публикации можно найти на <https://academia.edu> (только после бесплатной регистрации), <https://www.researchgate.net> (ниже под заголовком **Discover research**), а по языкознанию — ещё и на <https://ling.auf.net>.

Ресурсы СПбГУ Если вы знаете название статьи, которую хотите прочитать, её можно скачать через библиотеку СПбГУ. Для этого зайдите на <http://library.spbu.ru>. Пусть нужно найти статью “Movement Operations after Syntax”, опубликованную в журнале “Linguistic Inquiry” за 2001 год, т. 32, № 4. Тогда в разделе **Электронные журналы и книги A-Z** вводим в строке поиска название журнала...



Поиск по перечню электронных журналов и книг

✕

Поиск
?

Linguistic inquiry

Antisymmetry of Syntax (**Linguistic inquiry** monographs ; 25)

The Syntax of Adjectives: A Comparative Study (**Linguistic Inquiry** Monographs)

Argument Structure (**Linguistic inquiry** monographs ; 18)

Выбрав его, вы увидите результаты поиска. По ссылкам под названием журнала можно перейти на сайты, где находятся статьи из этого журнала. Слово *эмбарго* означает, что статьи последних лет будут недоступны.

Для скачивания статей понадобятся логин (**stXXXXXX**) и пароль от СПбГУ.

Результаты поиска: 1 - 50 из 72

Страница: 1 2 [Дальше](#)

1. Linguistic Inquiry

ISSN: 0024-3892, 1530-9150. **Linguistics**.

Поиск по изданию

■ [Доступ к полному тексту](#)

[JSTOR Archive Complimentary Collection \(COVID-19\)](#) 01.01.1970 - настоящее время (Время эмбарго на полный текст (в годах): 6)

[JSTOR Arts & Sciences VII Archive Collection](#) 01.01.1970 - настоящее время (Время эмбарго на полный текст (в годах): 6)

[Project MUSE - Standard Collection](#) 01.09.2000 Получить доступ к изданию через JSTOR Arts & Sciences VII Archive Collection. Откроется новое окно.

Выбрав один из сайтов, вы сможете просмотреть оглавление нужного номера журнала и уже в нём выбрать ту статью, которая вам нужна. Если статья доступна, где-нибудь на её странице будет ссылка типа Download PDF.



Search JSTOR [All Content](#)

[Advanced Search](#) [Browse](#) [Tools](#)



The MIT Press



Coverage: 1970-2014 (Vol. 1, No. 1 - Vol. 45, No. 4)
Published by: [The MIT Press](#)

All Issues

> 2010s

✓ 2000s

2009 (Vol. 40)

[No. 4 Fall, 2009 pp. 535-740](#)

[No. 3 Summer, 2009 pp. 367-533](#)

[No. 2 Spring, 2009 pp. 187-366](#)

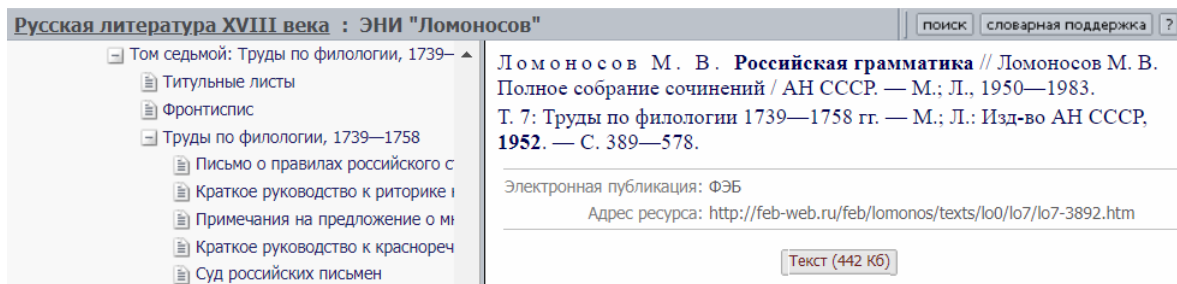
[No. 1 Winter, 2009 pp. 1-185](#)

Задание 71. Найдите через библиотеку СПбГУ статью «Movement Operations after Syntax» и напишите первое слово, напечатанное в этой статье на с. 573. ─

Задание 72. Воспользовавшись поиском библиотеки СПбГУ, назовите страницы, на которых опубликована статья «Ещё раз об энклитиках в „Слове о полку Игореве“» в №6 журнала «Вопросы языкознания» за 2007 год. Назовите последнее слово текста этой статьи (до списка литературы). ─

Старые книги Если книга старая, она уже не защищается авторским правом (copyright) и её можно законно скачать в Интернете. Есть сайты, где собраны такие книги.

Фундаментальная электронная библиотека <http://feb-web.ru> интересна в основном литературоведам, но там есть словари (раздел **Справочные разделы > Словари, энциклопедии** слева) и номера нескольких журналов (раздел **Наука**). Здесь же можно найти сочинения русских поэтов и писателей в научных изданиях (раздел **Литература и фольклор**); например, по ссылке **Ломоносов** в томе «Труды по филологии» можно найти текст «Российской грамматики».



Google Books <https://books.google.com> часто даёт просмотреть отдельные разделы новых книг, но старые там доступны целиком. Такие книги имеют ссылку **Читать**, а не **Предварительный просмотр**:

