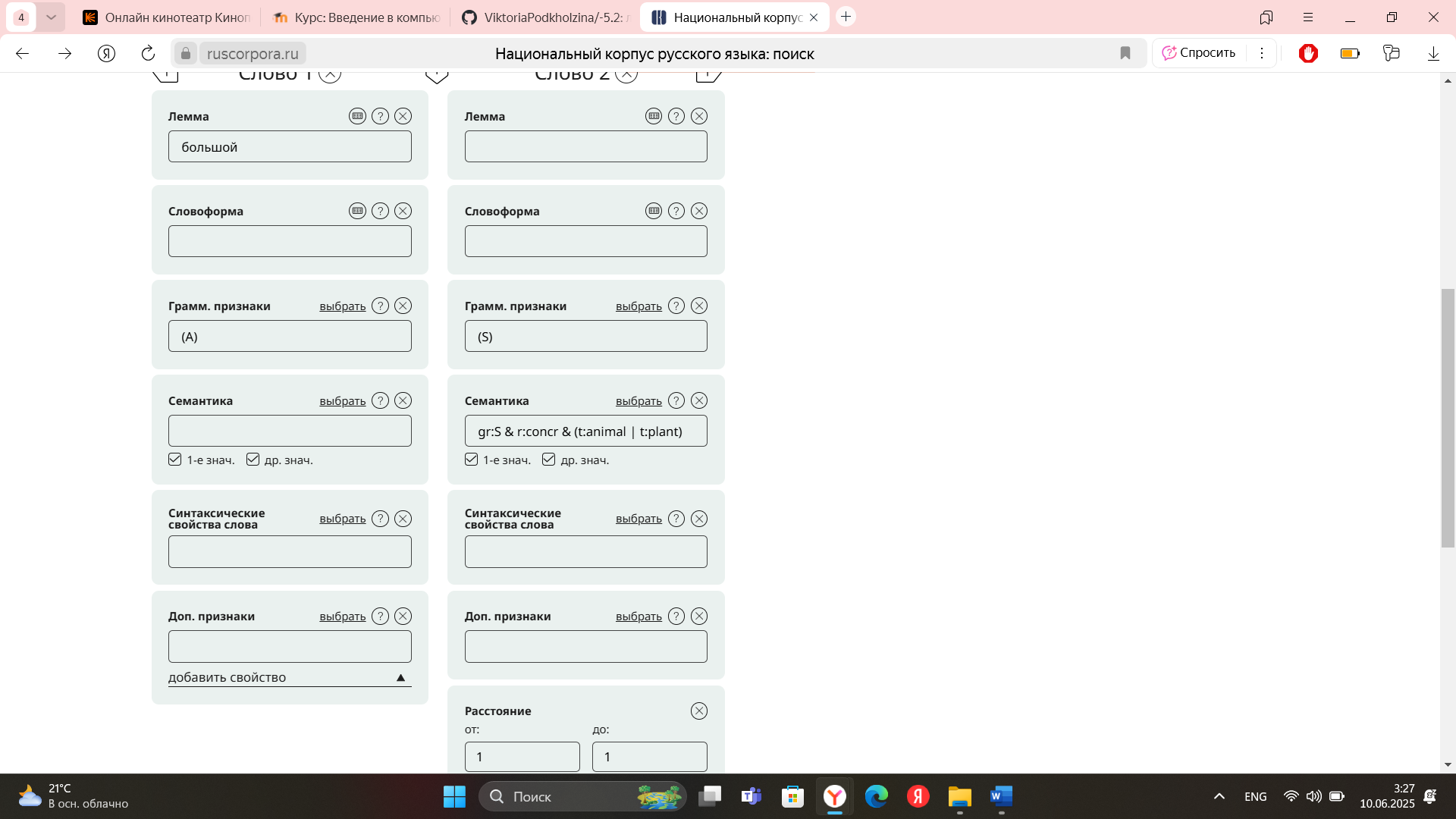
**Лабораторная №2**

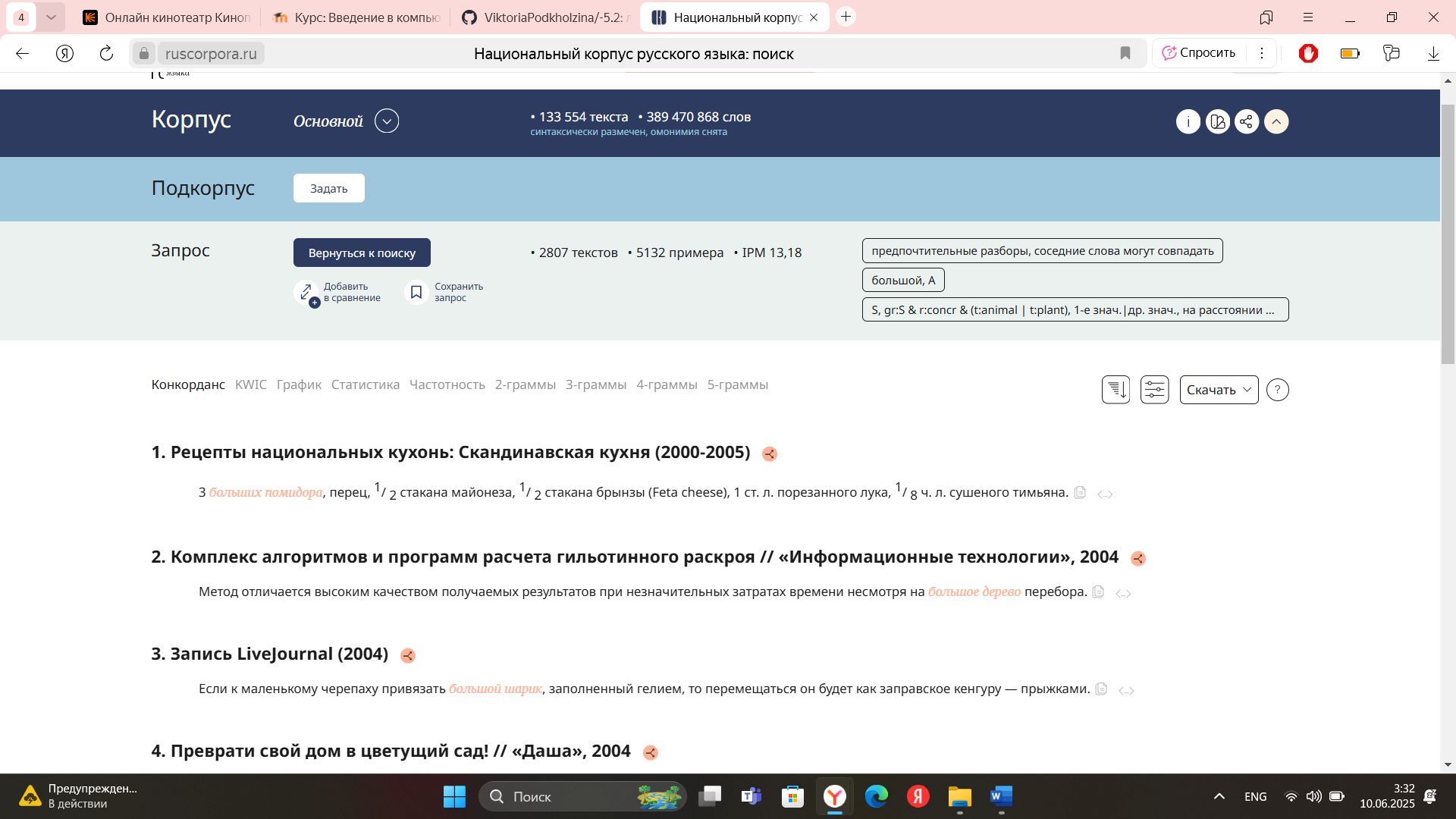
**Часть 1 (уровень 1)**

НКРЯ -<https://ruscorpora.ru/>

1. *Найдите в текстах основного корпуса все «большие растения и животные», т.е. первое слово должно быть прилагательным, обозначающим размер (большой), а второе слово – существительным, обозначающим растение или животное.*

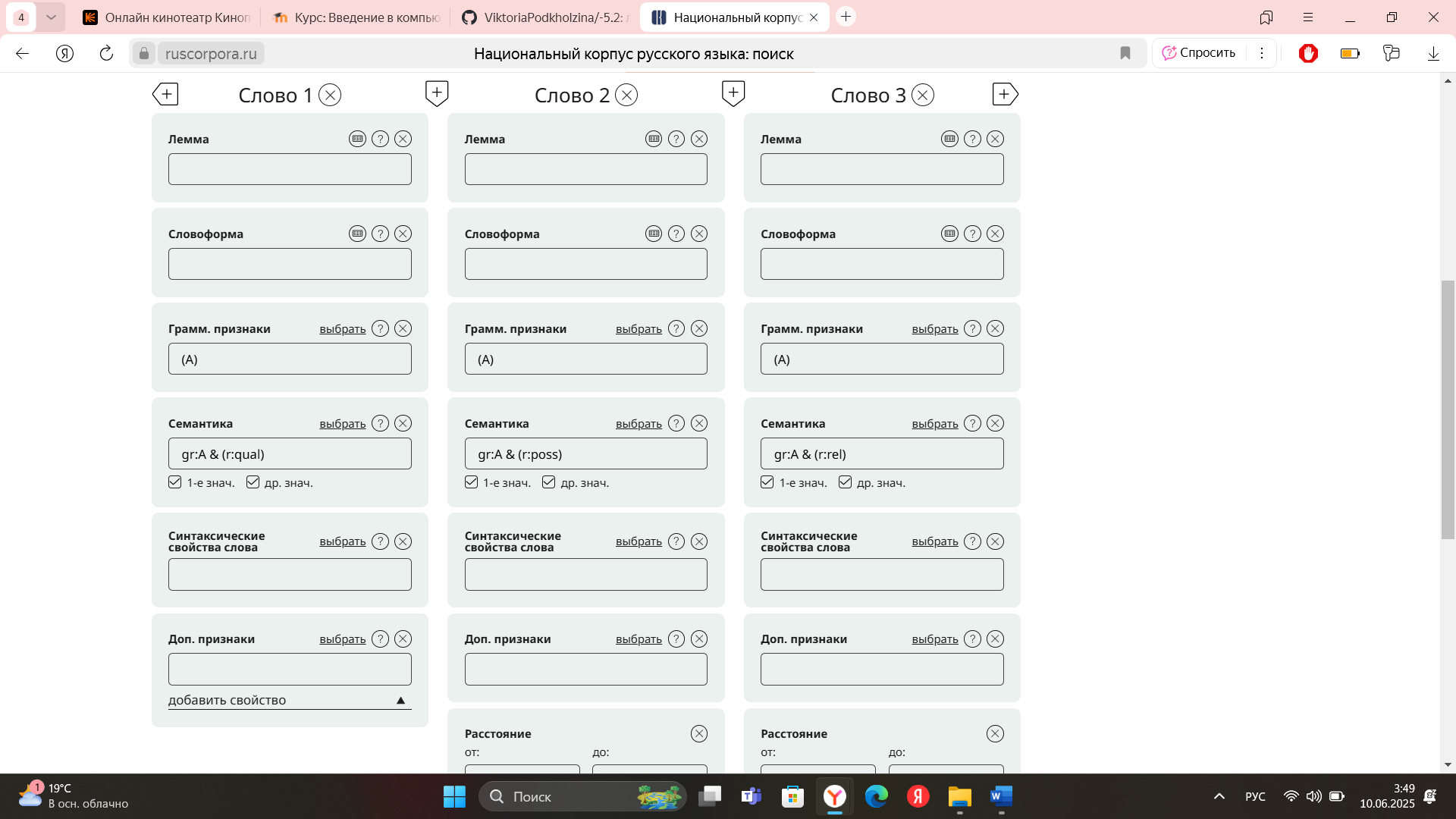
*Сколько таких вхождений?*

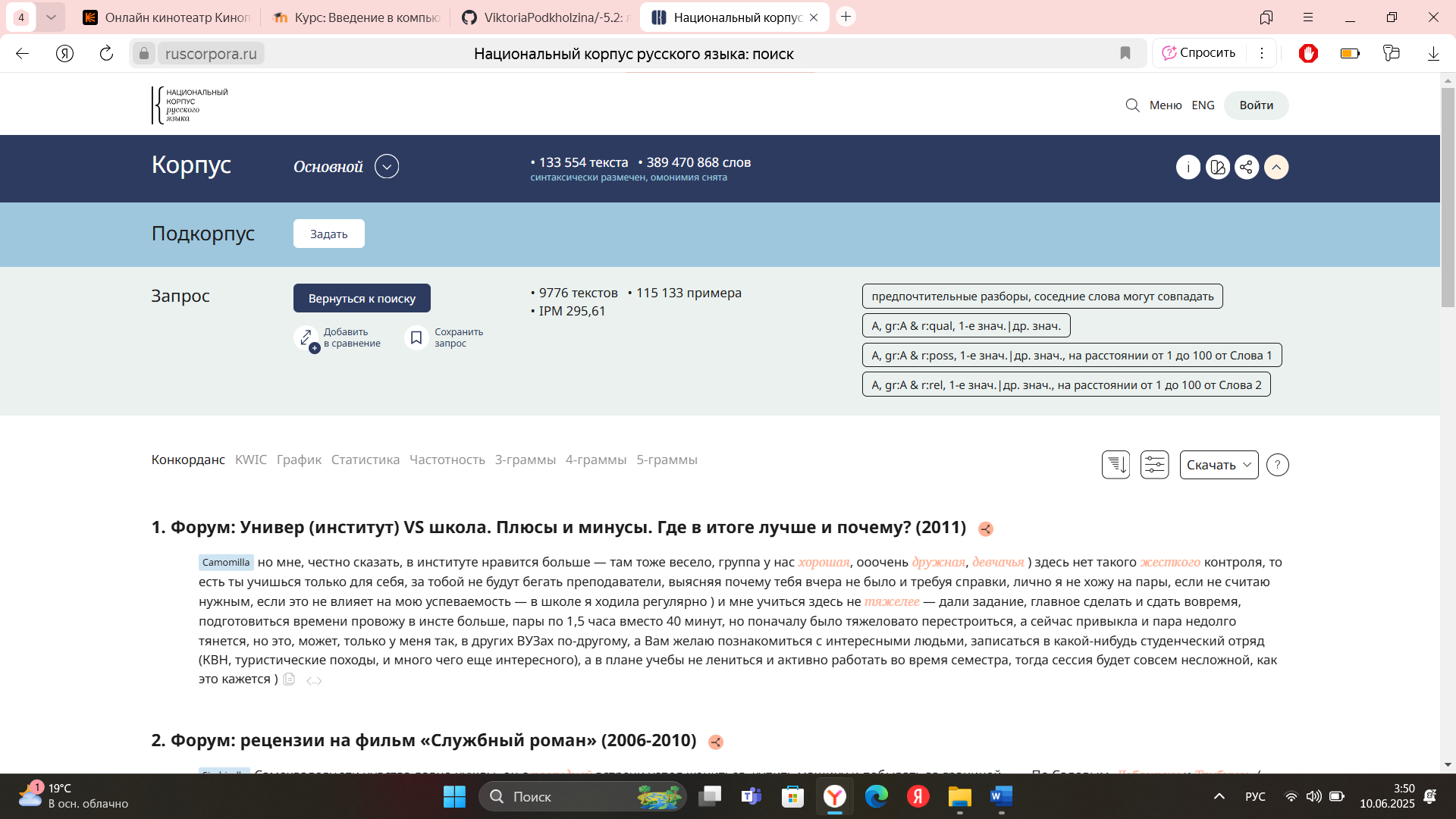
Запрос: 

Поисковая выдача: 

Ответ: всего вхождений – 5132 в 2807 текстах

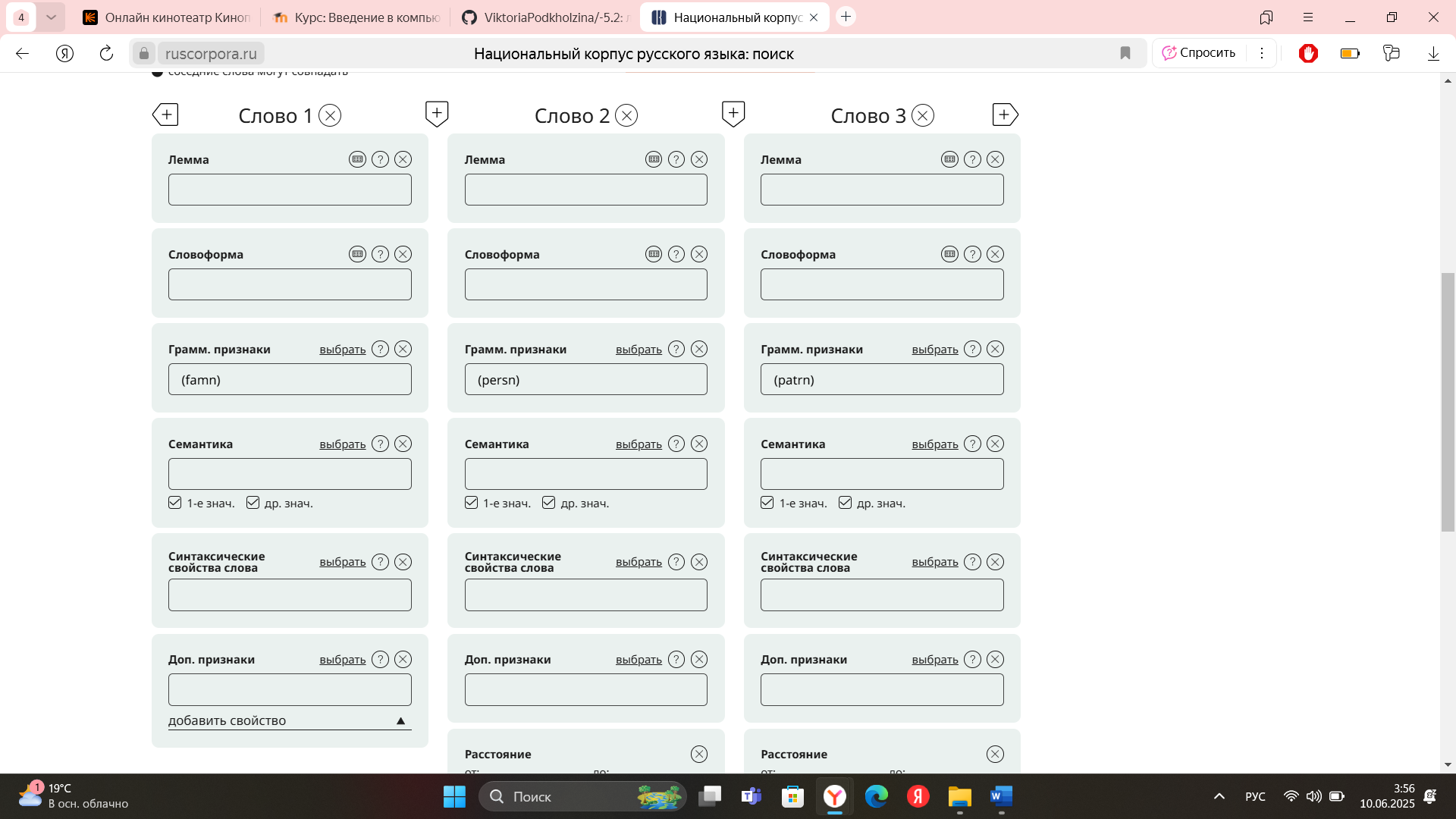
1. *В скольких предложениях есть три прилагательных, относящихся одно – к разряду качественных, другое – относительных, а третье – притяжательных?*

Запрос: 

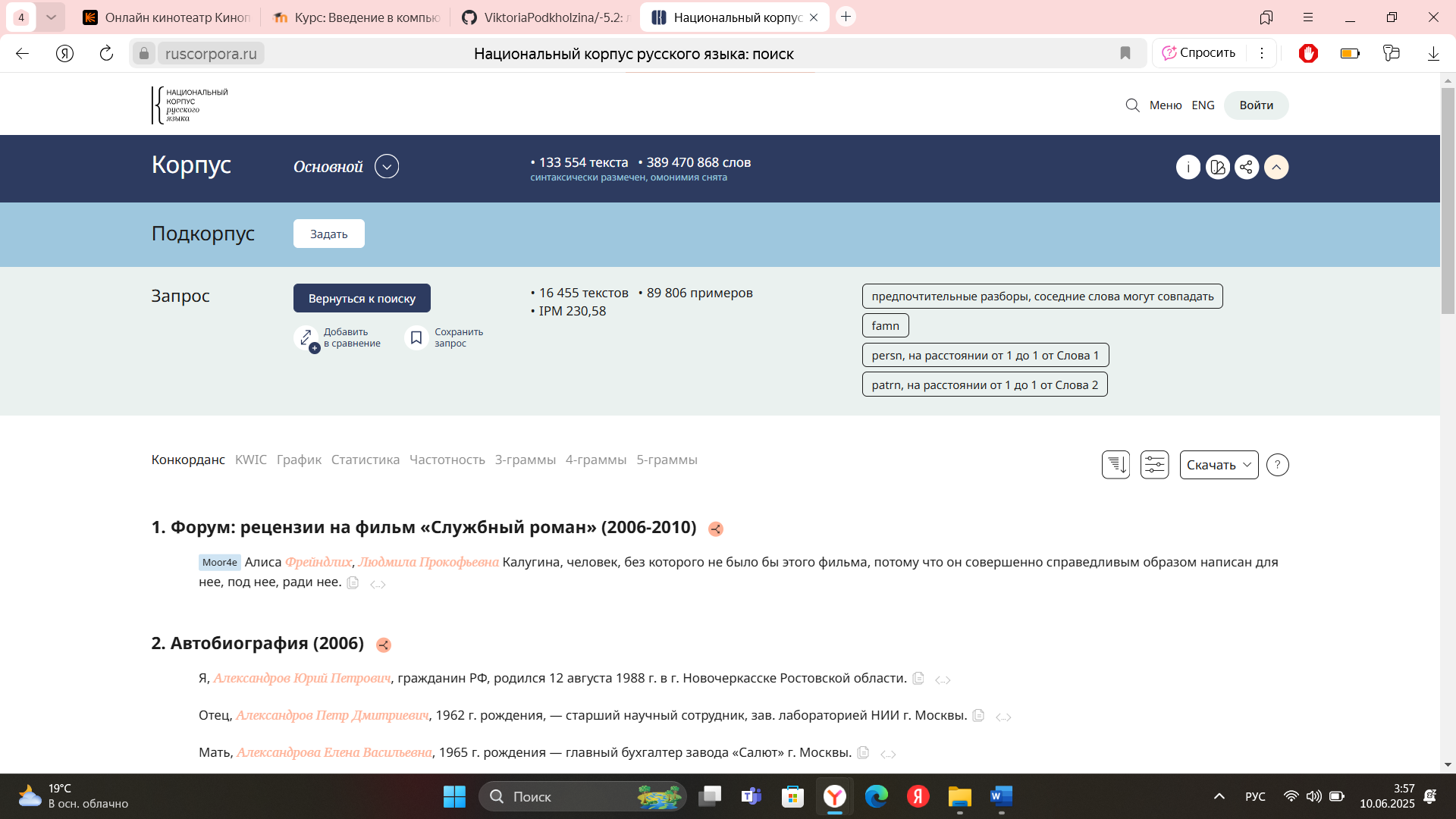
Поисковая выдача: 

Ответ: 115133 предложений

*В скольких предложениях после фамилии и имени нет отчества?*

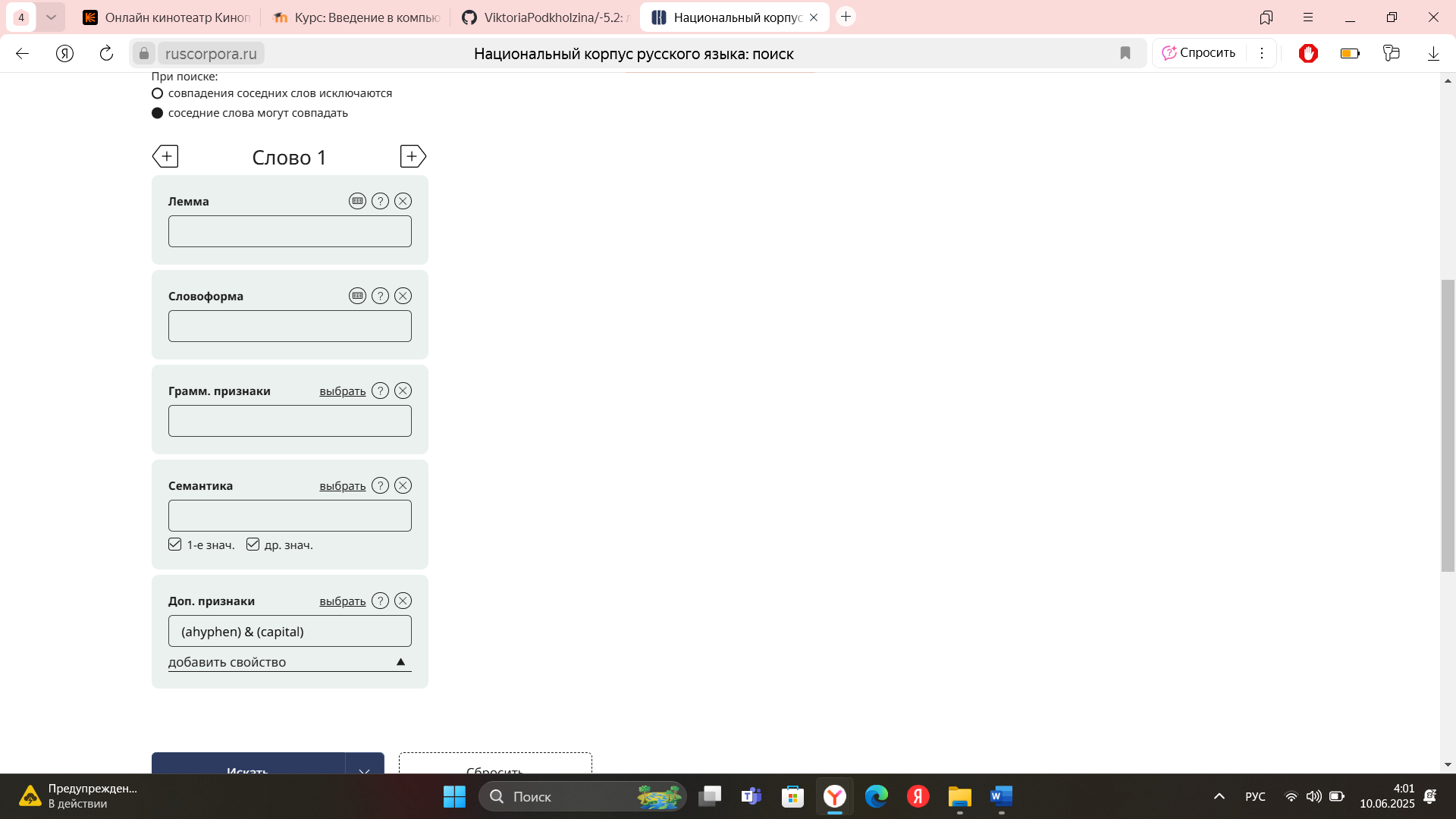
Запрос: **

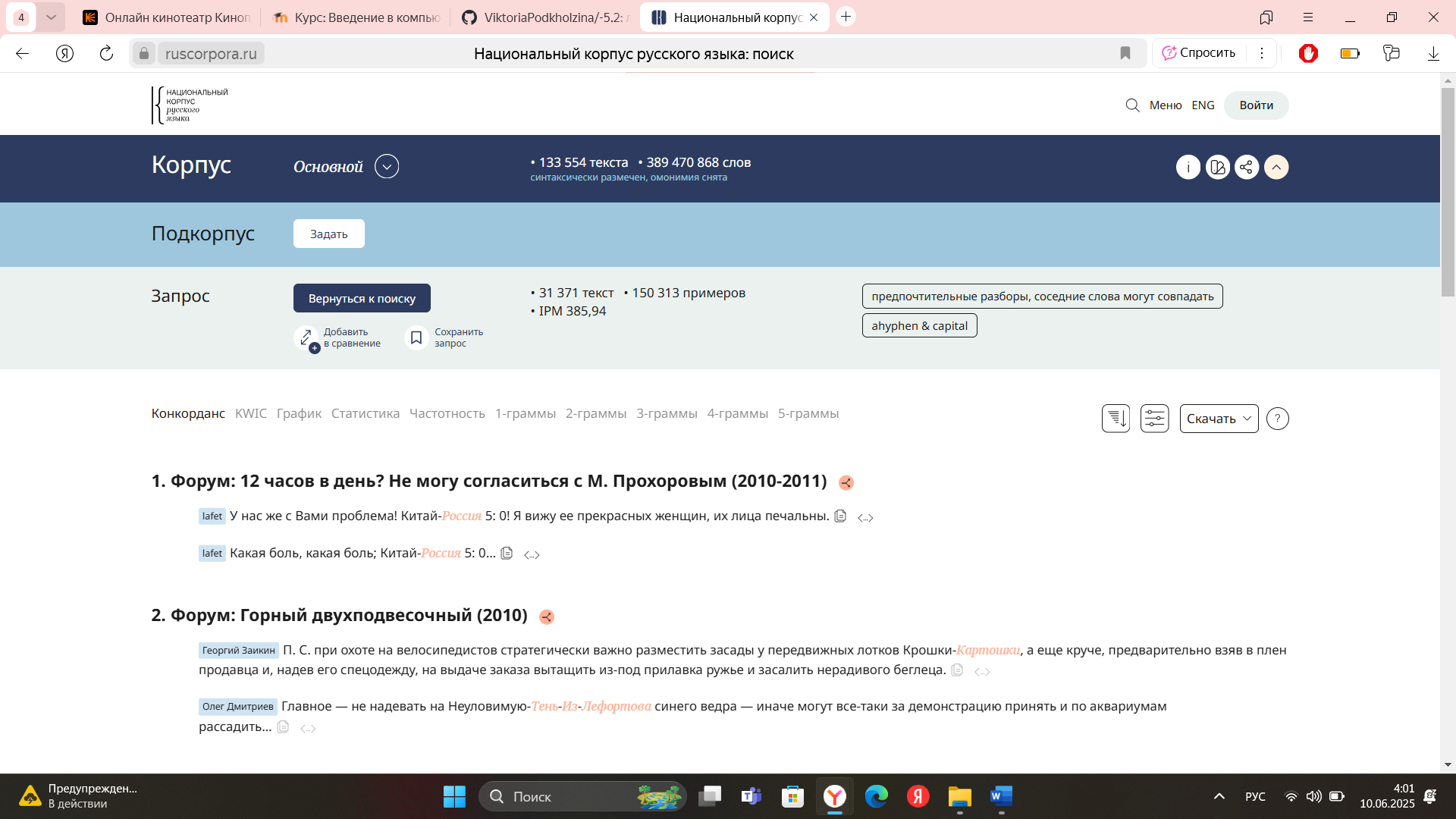
Поисковая выдача:

**

Ответ: 89806 примеров

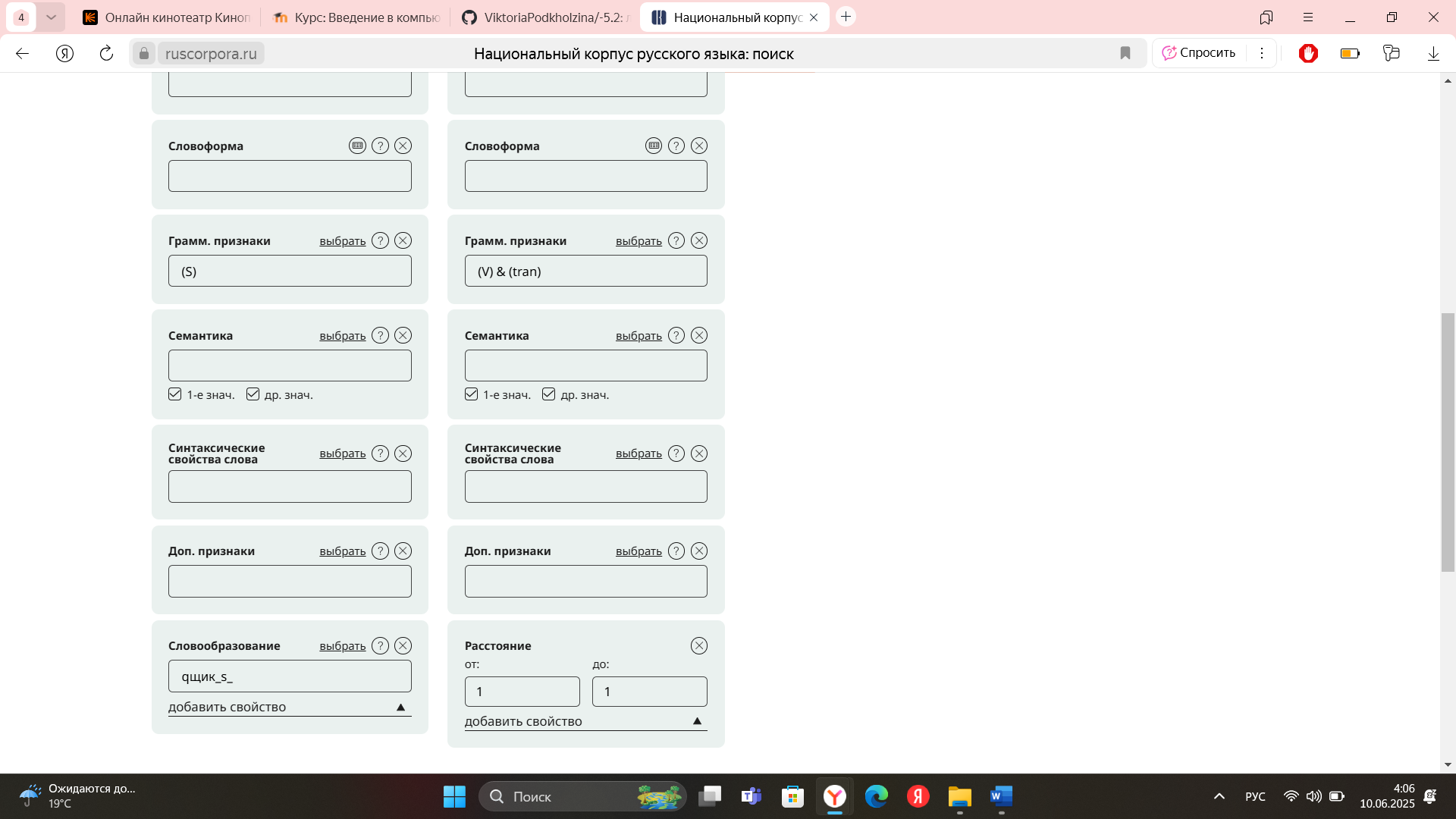
*4. В скольких предложениях слово с заглавной буквы стоит после дефиса?*

Запрос: 

Поисковая выдача: 

Ответ: 150 313 примеров

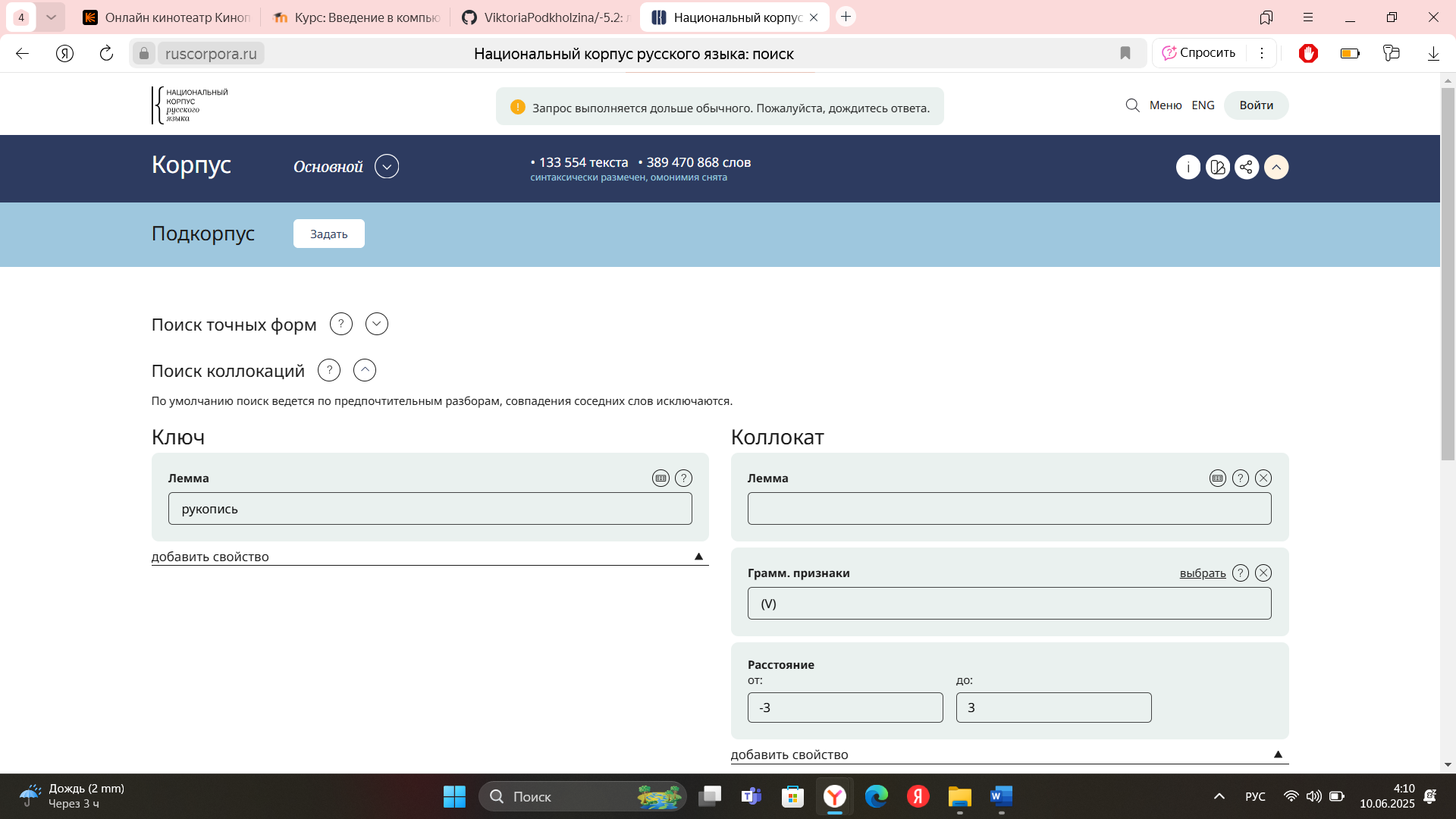
*Сколько предложений с существительным, оканчивающимся на суффикс -щик-, за которым следует переходный глагол?*

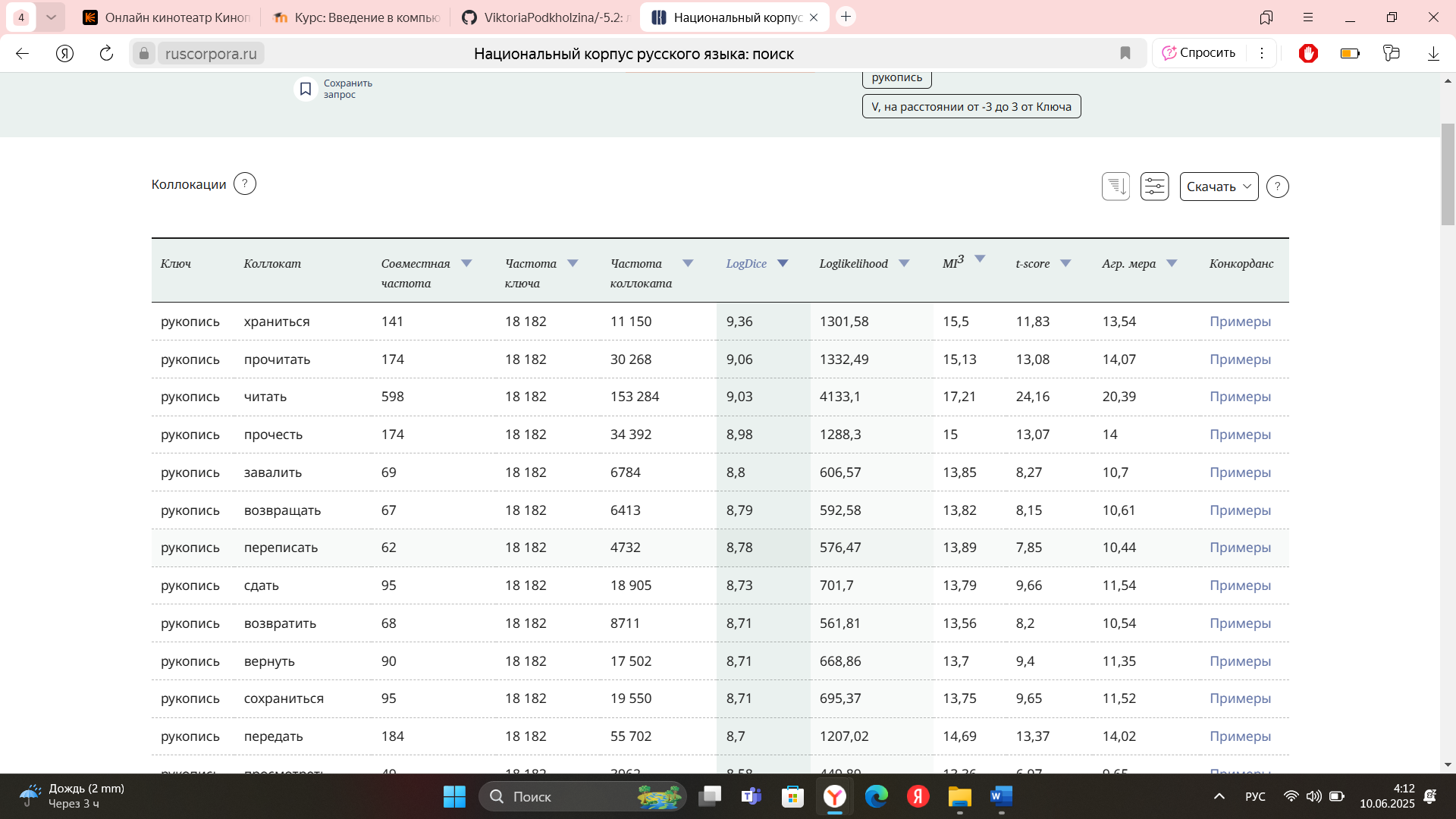
Запрос: 

Поисковая выдача: 

Ответ: 10 310 примеров

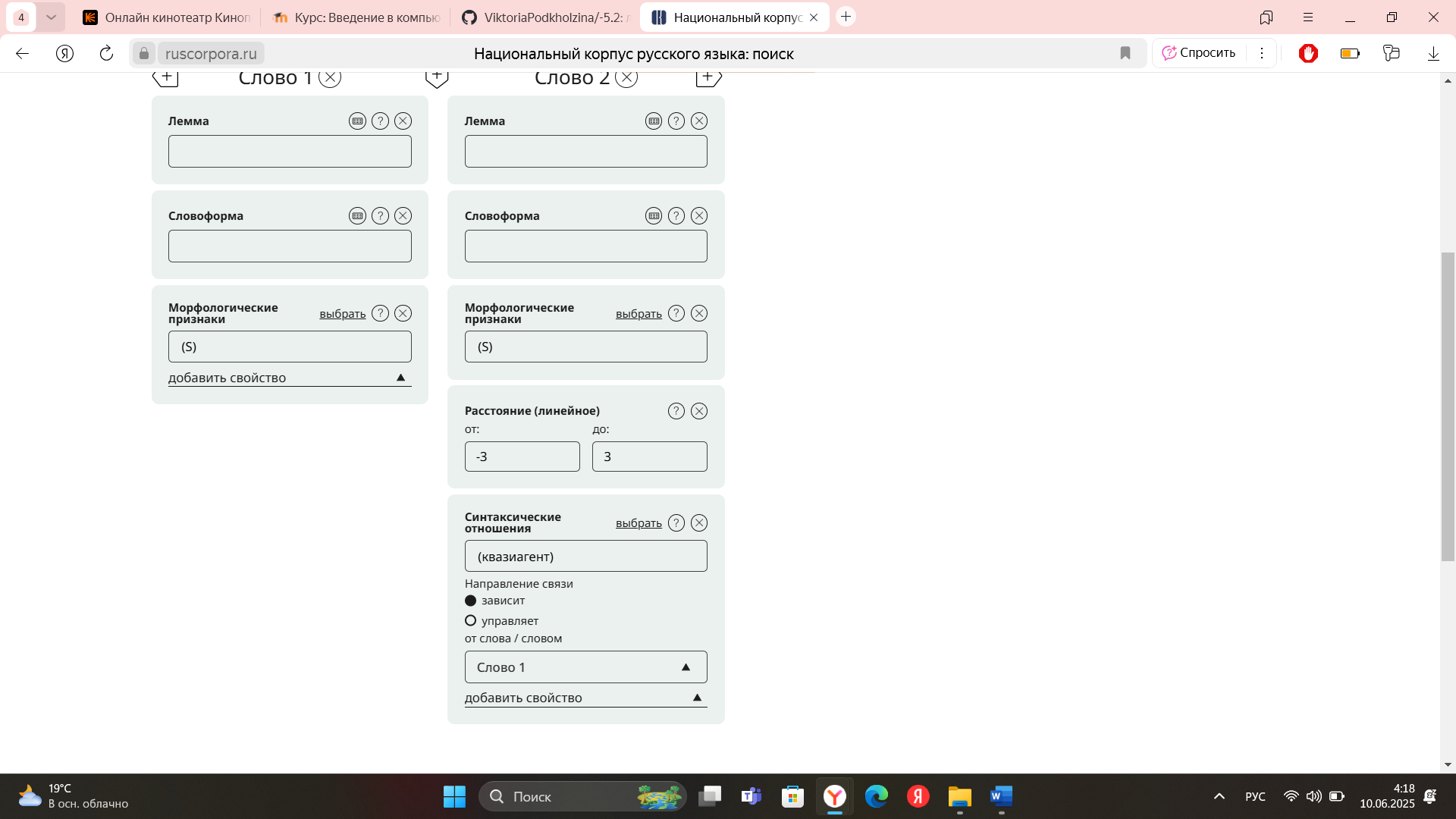
Какая самая частотная коллокация с существительным рукопись, если второе слово – глагол?

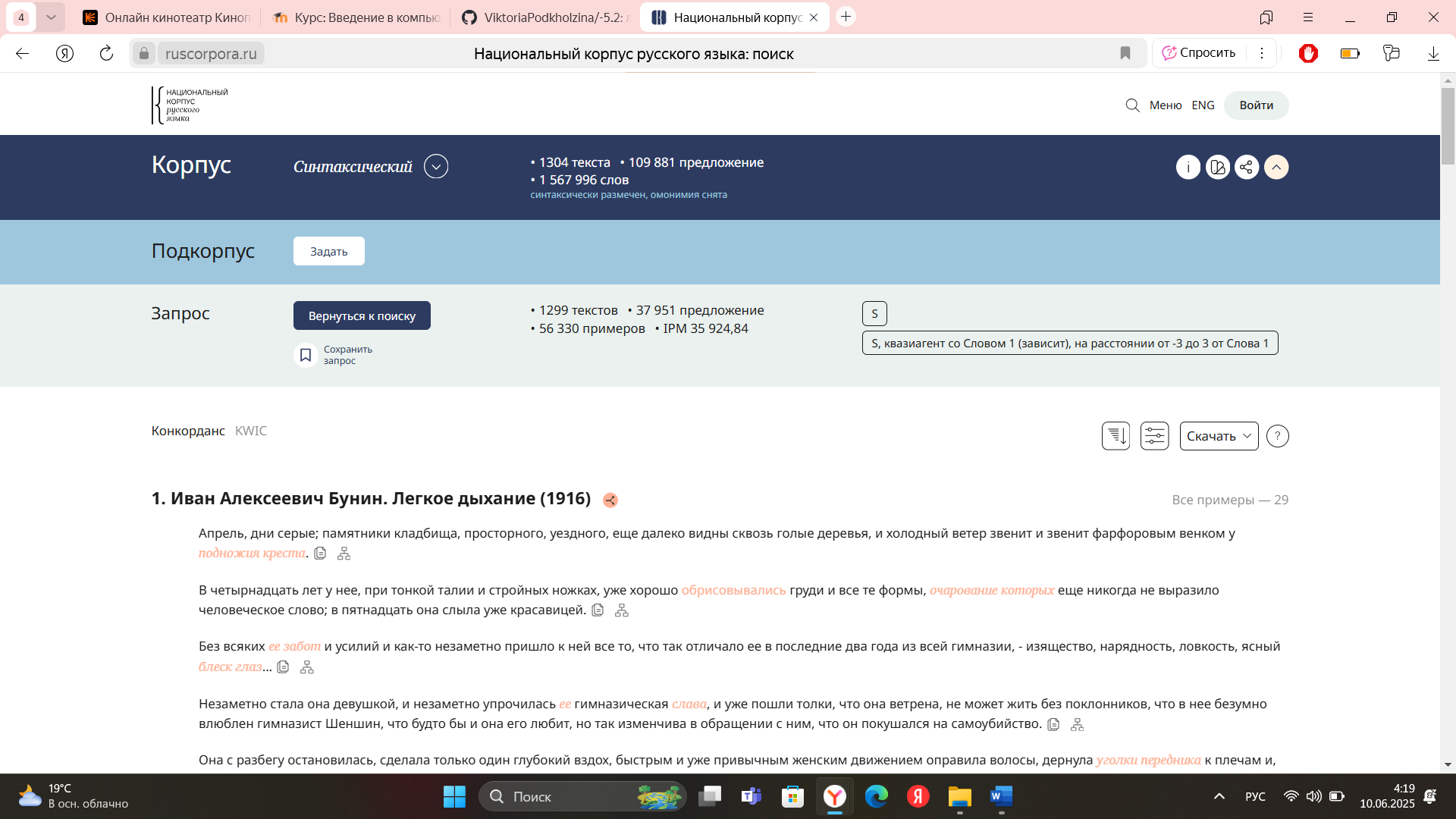
Запрос: 

Поисковая выдача: 

Ответ: храниться

Сколько в синтаксическом корпусе предложений, в которых оба слова – существительные, связанные квазиагентивным синтаксическим отношением и расстояние между словами от 1 до 2?

Запрос: 

Поисковая выдача: 

Ответ: 37951 предложений

**Часть 2 Базовый уровень (уровень 1)**

*Выберите один из англоязычных корпусов Марка Дэвиса. Опишите:*

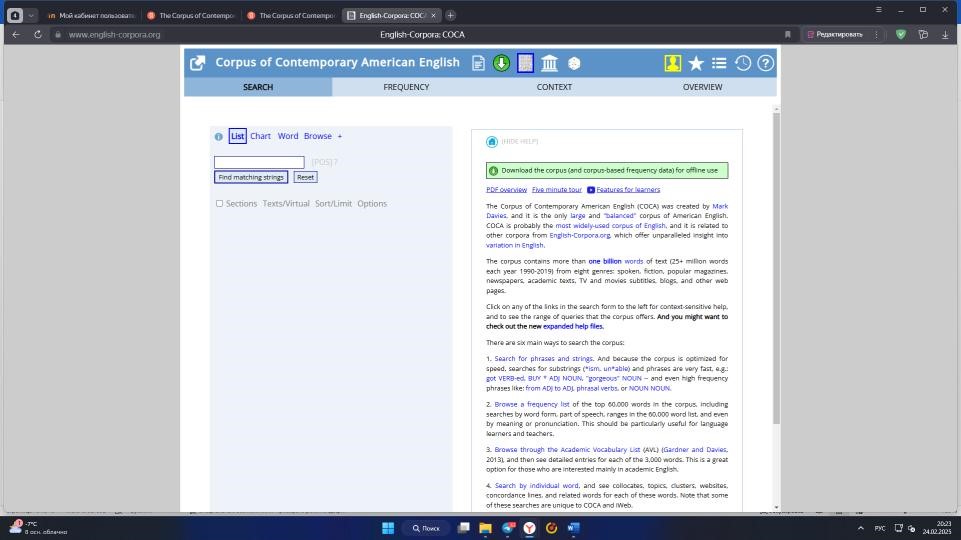
* *его функционал;*
* *последовательность действий при работе с данным ресурсом;*
* *сферы возможного применения;*
* *метасимволы/синтаксис, использующийся в выбранном корпусе.*

Для описания я выбрала корпус «The Corpus of Contemporary American English» (COCA) [Davies 2008-].

COCA представляет собой обширный корпус современного американского английского языка, включающий более 1 миллиарда слов, собранных из различных источников, таких как книги, журналы, газеты, телевидение и интернет.

Основной функционал включает:

* + Поиск и анализ лексических единиц и грамматических конструкций.
  + Сравнение частоты употребления слов и фраз в различных жанрах.
  + Доступ к контексту использования слов
  + Составление диаграмм



Последовательность действий:

1. Перейти на официальный сайт COCA ([https://www.english-corpora.org/Coca/)](https://www.english-corpora.org/Coca/)
2. Определить тип поиска, например провести простой поиск по ключевому слову или по частям речи или грамматическим конструкциям
3. Ввести запрос в строку поиска (слово или фразу)
4. При необходимости настроить параметры: фильтры по жанрам или временным периодам, минимальную частоту и т.д.
5. Изучить результаты поиска

Сферы возможного применения:

1. Лингвистика (исследование языка, анализ изменений в языке со временем, анализ употребления различных частей речи и грамматических конструкций)
2. Образование (использование в преподавании английского языка, особенно для изучения реального употребления языка, разработка учебных материалов на основе актуальных языковых данных)
3. NLP (использование корпуса для обучения моделей машинного обучения и алгоритмов обработки естественного языка) Метасимволы/синтаксис:
   * Обычные слова вводятся без каких-либо символов. Например, dog найдет все случаи употребления слова "dog".
   * Использование кавычек позволяет искать точные фразы. Например, "New York" найдет все случаи, где встречается именно эта фраза.
   * (\*) используется как подстановочный символ для одного или нескольких символов. Например, teach\* найдет слова, начинающиеся на "teach", такие как "teacher", "teaching" и т.д.
   * (?) заменяет один символ (например, "wom?n" найдет как "woman", так и "women")
   * [ ] позволяет указать диапазон символов (например, "gr[aeiou]p" найдет "grap", "grep", "grop" и т.д.).
   * (#) используется для обозначения количества слов между заданными словами (например, "word1 # word2" найдет предложения с одним или несколькими словами между "word1" и "word2").
   * (+) используется для указания на обязательные слова в запросе. Например, +love +hate вернет результаты, содержащие оба слова.
   * (-) используется для исключения слов из результатов. Например, dog -cat найдет все случаи слова "dog", но исключит те, которые содержат "cat".
   * Скобки () позволяют группировать условия. Например, (dog OR cat) вернет результаты, содержащие либо "dog", либо "cat".
   * Использование двоеточия для указания диапазона. Например, 2000:2020 может использоваться для поиска данных в определенные годы.
   * COCA позволяет выполнять запросы с использованием тегов частей речи, таких как NOUN, VERB, ADJ и т.д. Например, [NOUN] будет искать существительные.

# Часть 3 Базовый уровень (уровень 1)

1. *Выберите два слова, находящихся в одном семантическом поле (парадигме, где все компоненты имеют хотя бы один общий семантический признак, например, движение: бег, плавание, ходьба…): гипотетически лингвоспецифичное и неспецифичное. Используя вышеперечисленные параллельные корпусы*

*(Параллельные корпуса НКРЯ, OPUS, Linguee) и меры разброса, подтвердите или опровергните свои гипотезы относительно их лингвоспецифичности или ее отсутствия.*

1. *Приведите минимум 3 примера аналогичных ресурсов с параллельными текстами.*

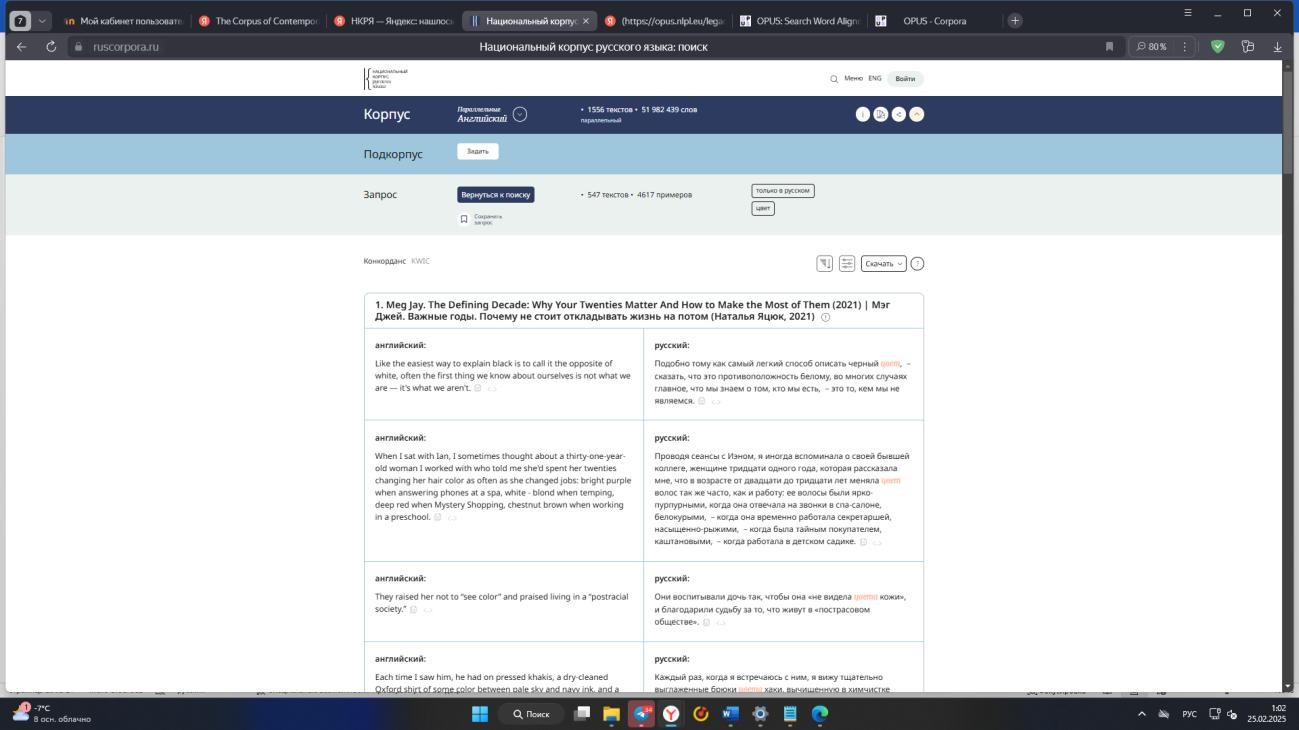
Выберу 2 слова с общим семантическим признаком – физические свойства:

Лингвоспецифичное - упругость

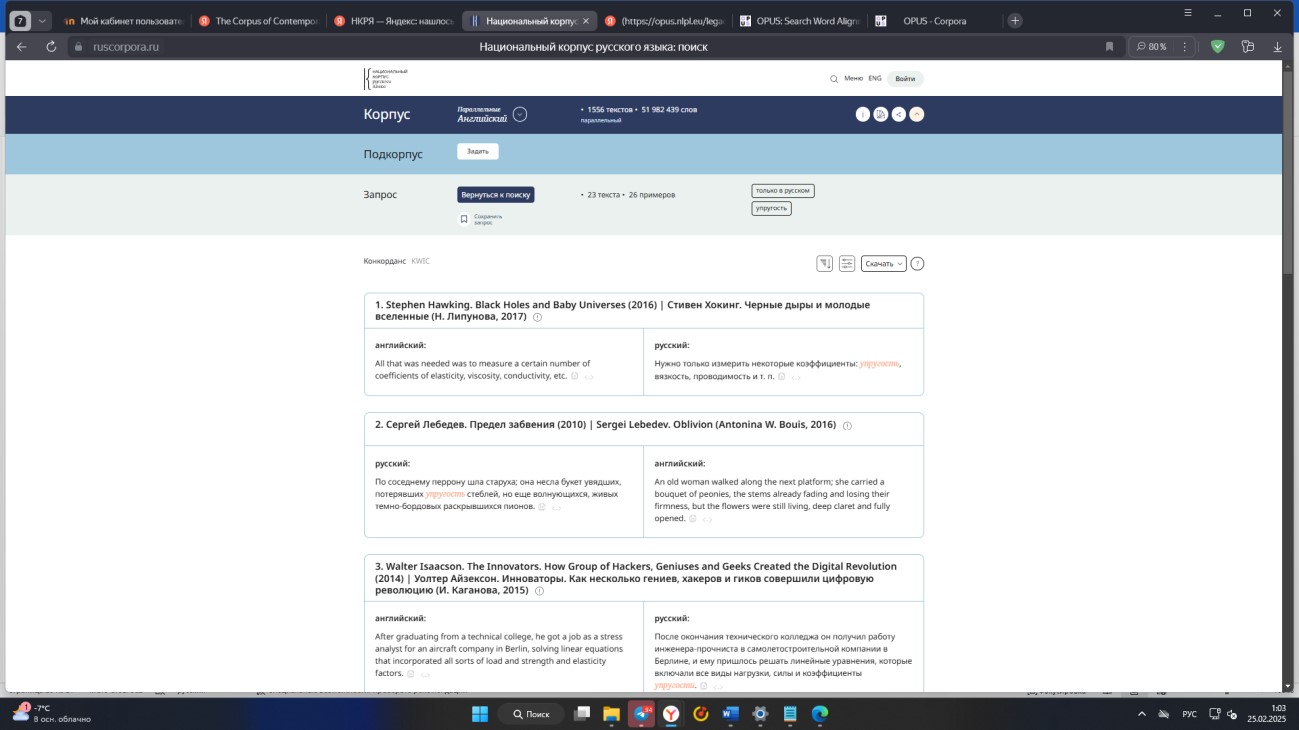
Неспецефичное – цвет

1. НКРЯ

Запрос «цвет» вывел 4617 примеров:

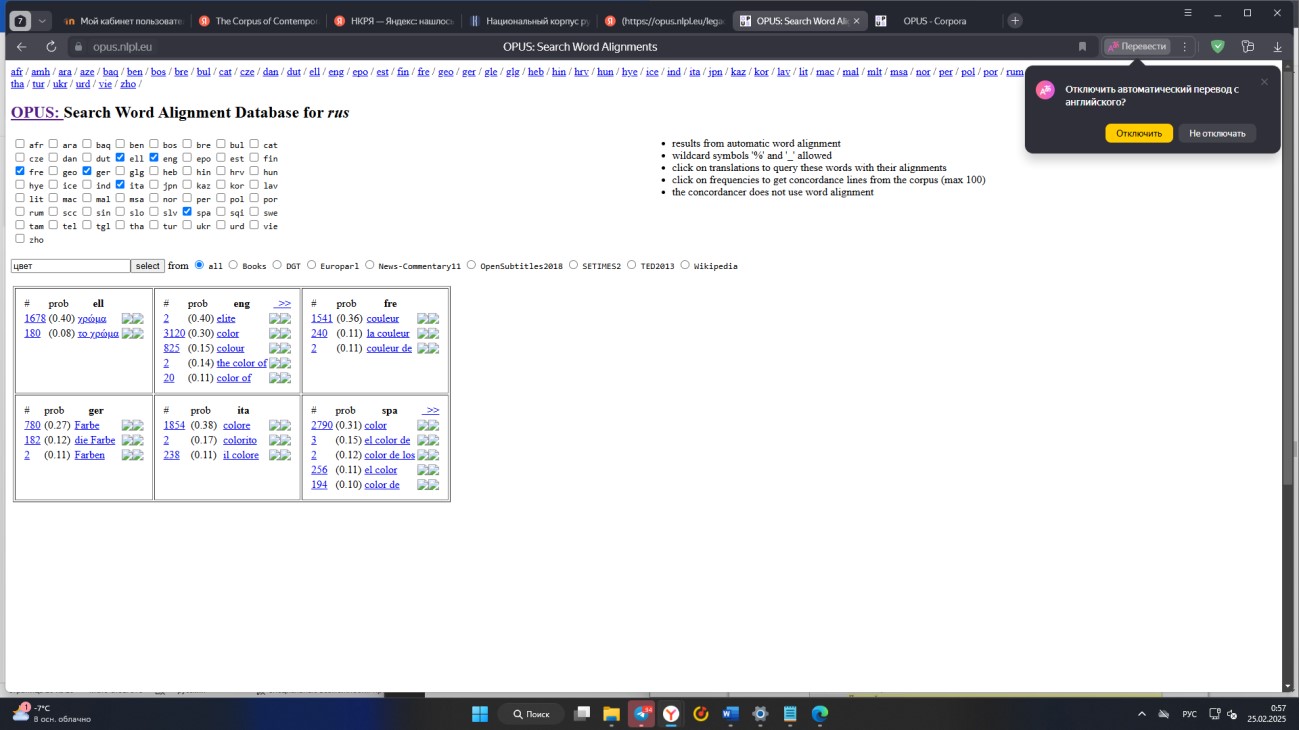


Запрос «упругость» вывел всего 26 примеров:

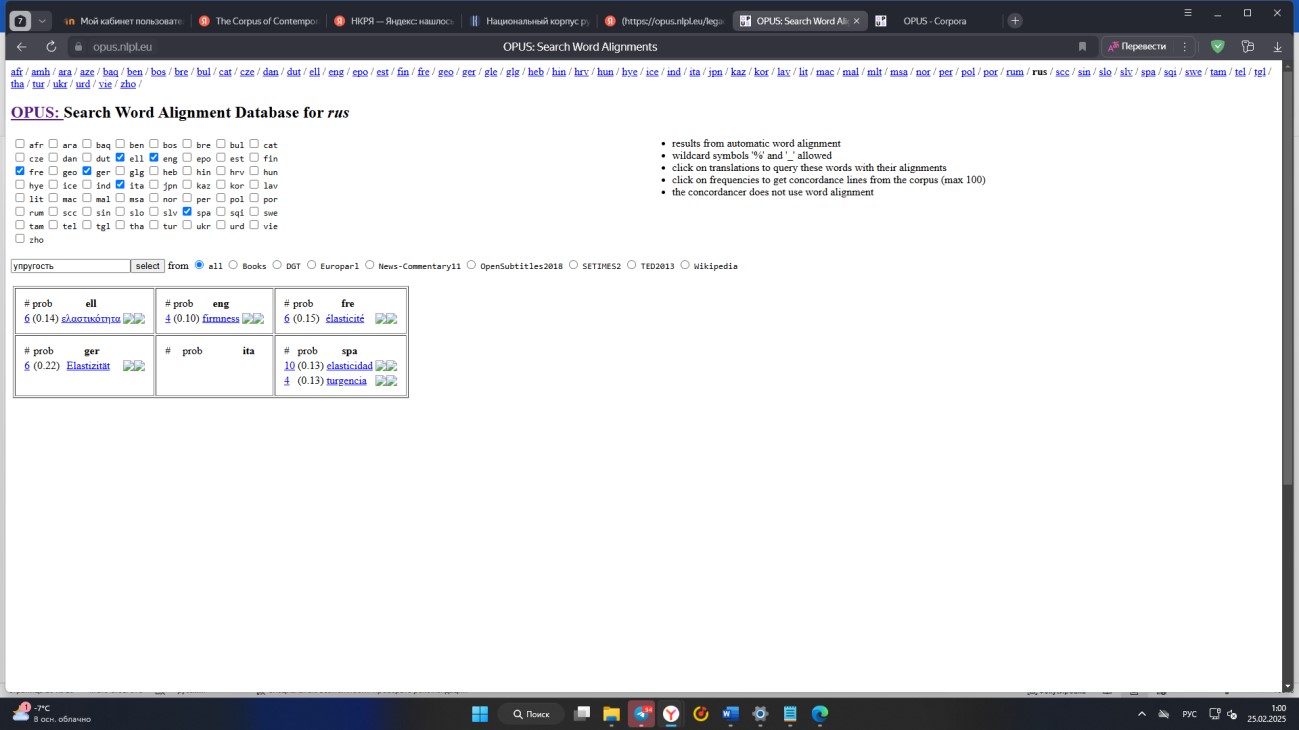


1. OPUS

Запрос «цвет» вывел множество различных переводов:

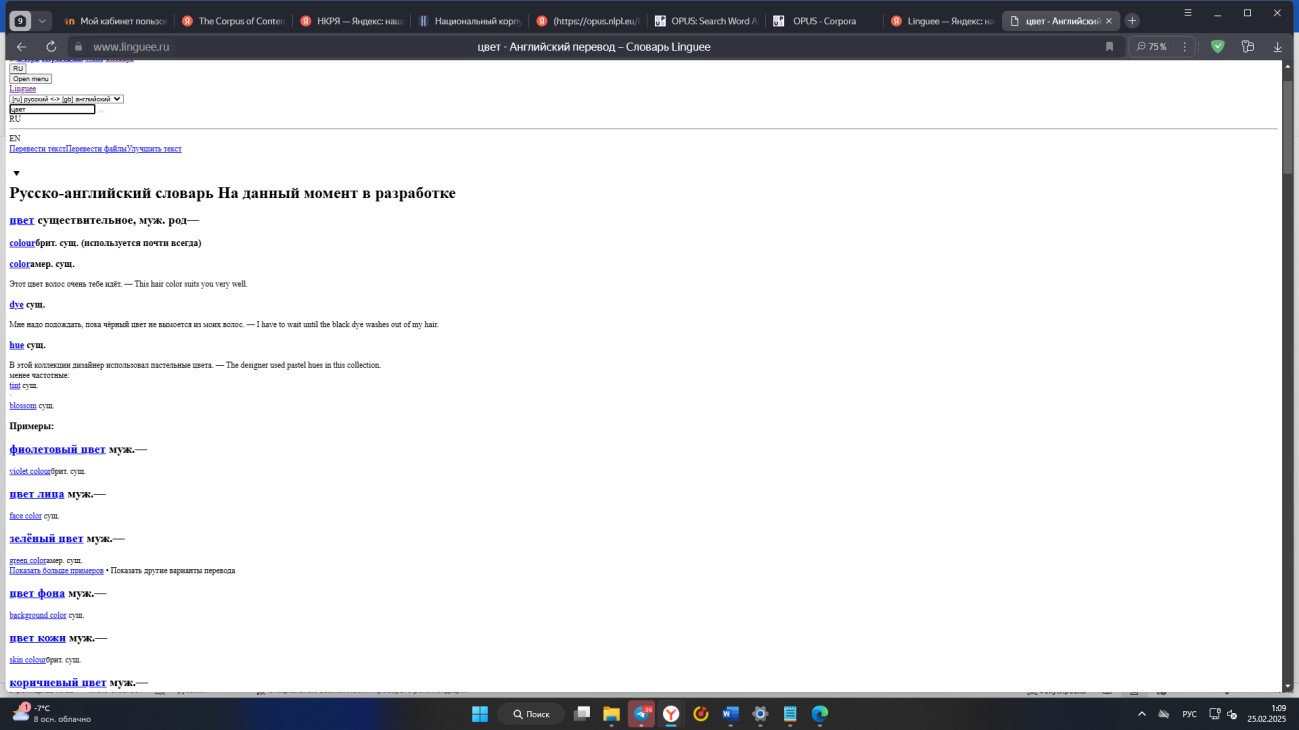


Запрос «упругость» вывел намного меньше переводов:

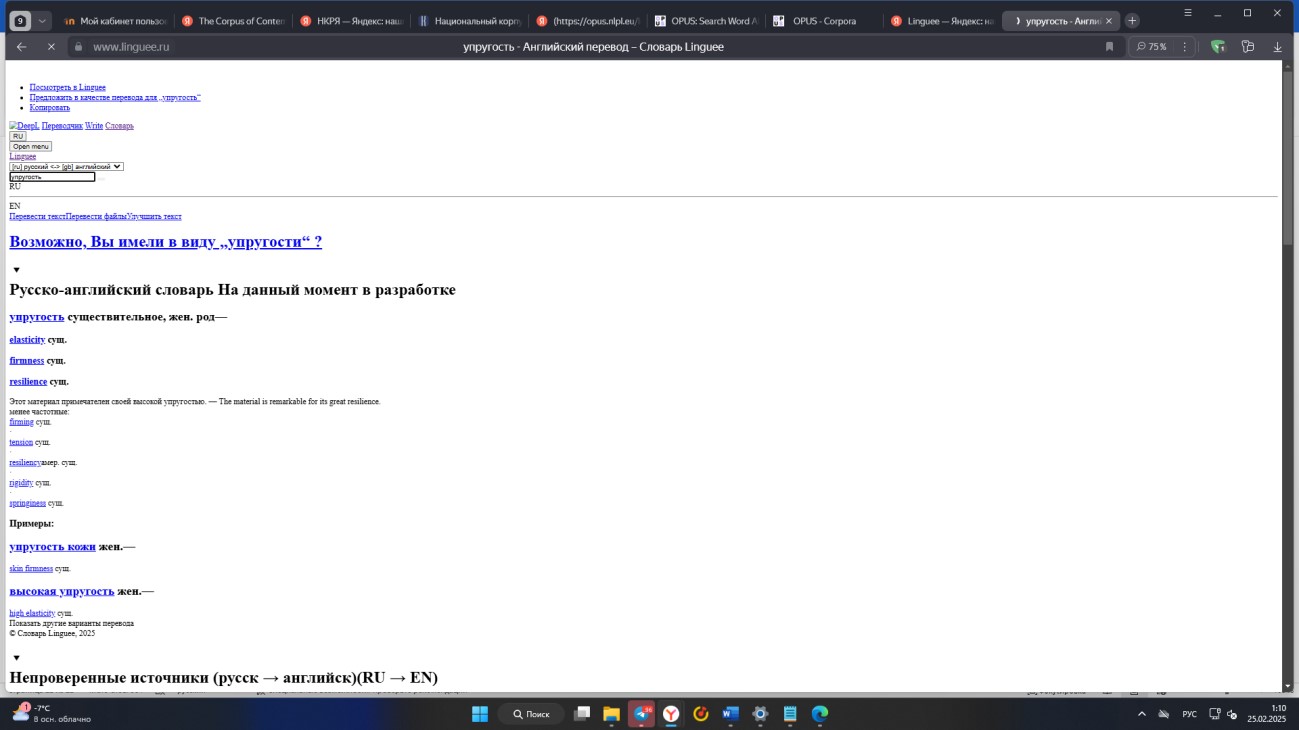


1. Linguee

Запрос «цвет» вывел множество различных словосочетаний, в которых употребляется данное слово:



Запрос «упругость» вывел всего 2 словосочетания:



Вывод: линейные корпусы дали понять, что моя гипотеза оправдалась и действительно, даже слова, находящиеся в одной семантической группе, могут быть более или менее специфичными

Примеры аналогичных ресурсов с параллельными текстами:

Project Gutenberg ([https://www.gutenberg.org/)](https://www.gutenberg.org/)

Open Library ([https://openlibrary.org/)](https://openlibrary.org/)

Bilingual Books ([https://www.bilingualbooks.com/)](https://www.bilingualbooks.com/)