**Лабораторная работа №3 корпусные исследования**

**Кругликова Елена Владиславовна, Л-14**

1. Изучите основную информацию о национальном корпусе русского языка (НКРЯ) на сайте http://www.ruscorpora.ru , раздел FAQ <https://ruscorpora.ru/page/faq/> , сайт <https://studiorum.ruscorpora.ru> и ответьте на следующие вопросы:- Кто является создателем НКРЯ?

Всё началось с Центра лингвистической документации. Так назывался кружок московских лингвистов под руководством В. А. Плунгяна и М. А. Даниэля, проводивший цикл семинаров в Центре непрерывного математического образования по приглашению руководителей центра. Инициатива, как и во многих других лингвистических начинаниях, исходила от математиков - И. С. Красильщика и С. К. Ландо. Идея создания кружка состояла в том, чтобы объединить разрозненные усилия лингвистов и программистов по созданию исследовательских инструментов нового поколения: баз данных, программ автоматического поиска, интеллектуальных систем, а также корпусов текстов. Первый семинар состоялся 13 октября 2000 года, и на нём уже обсуждались основные темы, которые стали лейтмотивом дальней- шей деятельности кружка. Говорилось о том, что несмотря на давнее и постоянное внимание отечественных лингвистов к проблеме компьютеризации и информатизации своей науки - прежде всего, к удобному хранению языковых данных в компьютерном виде - в этой области имеется до обидного мало результатов, доступных всему лингвистическому сообществу.

В отечественной прикладной лингвистике зияла парадоксальная брешь - по сути, не было общедоступного представительного и аннотированного (размеченного) корпуса русского языка, с которым можно было работать лингвисту. Отечественная корпусная лингвистика началась как минимум с 1980-х годов, с работ А. П. Ершова и В. М. Андрющенко, посвящённых тому, «каким должен быть» машинный фонд русского языка. Тогда же в Институте русского языка имени В. В. Виноградова РАН был создан Отдел машинного фонда, занимавшийся, в частности, переводом русских литературных текстов в электронную форму. К сожалению, технических и организационных возможностей для того, чтобы создать полноценный машинный фонд текстов русского языка, в тот период не хватило.

У В. А. Плунгяна, М. А. Даниэля, А. А. Егорушкина и некоторых других (обсуждавших эту идею в первые месяцы 2001 г.) возникали варианты названия вроде «Стандартный корпус русского языка», «Представительный корпус русского языка» (под «представительностью» мы тогда имели в ви- ду довольно субъективную «нормативность»), «Русский стандарт» (в чём была изрядная доля шутки - так называлась активно рекла- мировавшаяся в то время водка, а на Воздвиженке около Институ- та языкознания висел рекламный щит этого напитка, который нас вО многом вдохновлял: «РУССКИЙ СТАНДАРТ: СДЕЛАТЬ НЕ- ВОЗМОЖНОЕ»). Технически-программистскую поддержку корпус получил со стороны компании «Яндекс», ранее уже участвовавшей в лингвистических проектах и заинтересованной в корпусе как в важном средстве совершенствования целого ряда программ автоматической обработки текста; тогда же состоялись первые встречи участников ЦЛД с техническим директором компании И. В. Сегаловичем и программистом В. А. Титовым, которые до сих пор являются ключевыми участниками работ по Корпусу.

( http://www.inforeg.ru/electron/concord/concord.htm ). Сотрудничество облегчало то, что созданными «Яндексом» при участии сотрудников Лаборатории компьютерной лингвистики ИППИ РАН (программа Mystem); им также был разработан шаблон для правки разметки в текстовом редакторе и интерфейс поиска, начались работы по снятию морфологической омонимии в небольшом - тогда ещё даже менее 1 млн. словоупотреблений - корпусе, состоявшем из художественных текстов второй половины XX в. (всего несколько авторов: Аксёнов, И. Грекова, Трифонов, Довлатов, Татьяна Толстая). Снятие омонимии (тогда оно было полностью ручным) финансировал «Яндекс», а работали студенты и аспиранты МГУ и РГГУ. В сентябре 2001 г. мы с И. В. Сегаловичем и А. Е. Поляковым сделали доклад о корпусе на очередном после летних каникул семинаре ЦЛД -- это был уже не широковещательный прожект, а отчёт о действительно проделанной работе. Данный этап деятельности был отражён в моей статье, которую тогда же и отдали в печать, но к моменту выхода она, что достаточно показательно для истории нашего Корпуса, успела сильно устареть.

Теперь о материальной стороне дела: бюджете. Корпус получил поддержку по ряду грантов. Первым, как я уже сказал, поддержку нашему начинанию оказал «Яндекс»: деньги были потрачены на разметку «пилотного» корпуса с помощью программы И. В. Сегаловича и на снятие в нем омонимии. Потом наступил финансовый перерыв - в это время мы решили попробовать перейти на другую программу морфологического анализа, которая базировалась на грамматическом словаре А. А. Зализняка, - Dialing, создававшейся А. В. Сокирко и Д. В. Панкратовым (изначально - для системы, которая делалась под руководством Н. Н. Леонтьевой). Но всякую, даже очень хорошую программу, надо «дотягивать» и приспосаб- ливать - и мы гордимся, что, совершенствуя снятие омонимии, так сказать, в новых условиях, мы работали бок о бок с этими программистами. Тут нам помог молодежный грант Президиума РАН (ведь омонимию снимали в основном студенты и аспиранты!). Интересы создателей Корпуса быстро вышли за пределы ком- пактного корпуса текстов «хороших писателей».

Корпус - это прежде всего люди. Энтузиазм, идеи (разной степени «безумности») и просто человеческое участие и помощь. Число людей, имеющих отношение к этому проекту, непрерывно растет. С конца 2002 г. корпус со снятой омонимией стал доступен для поиска - пока ещё под паролем по адресу www.ruscorроrа.ru, на сервере компании «Яндекс». Интересно, что буквально с того вре- мени, как «Яндекс» зарегистрировал это доменное имя, к нему было приковано внимание деятелей Интернета - моментально в службу поддержки стал приходить поток писем: «что это?», «что это за новый проект Яндекса?» И в дальнейшем Национальный корпус привлекал интерес программистов, компьютерщиков и всех, кто интересовался развитием Рунета. Уже в первые месяцы после официального открытия Корпус стал предметом рецензий в профессиональной компьютерной среде, для которой теоретическая лингвистика представляла лишь косвенный интерес, на таких сайтах, как «Вебинформ». Информация о сайте, несмотря на о сутствие открытого доступа, распространилась среди лингвистов, причём не только в России, достаточно быстро; так, уже весной 2003 года он регулярно использовался в преподавании русского языка в Карловом университете Праги (в чём принял некоторое участие и автор этих строк, посетивший тогда традиционную Школу Матезиуса и ставший свидетелем огромного интереса к нашему проекту как со стороны математических лингвистов, создателей Чешского национального корпуса, таких, как В. Петкевич или Я. Гайич, так и русистов - преподавателей Восточнославянской кафедры философского факультета). В проекте участвуют специалисты Института русского языка им. В. В. Виноградова РАН [[ИРЯ РАН](http://www.ruslang.ru/" \t "https://ruscorpora.ru/page/corpora-about/_blank)], Национального исследовательского института «Высшая школа экономики» [[ВШЭ](http://www.hse.ru/" \t "https://ruscorpora.ru/page/corpora-about/_blank)], Института проблем передачи информации РАН [[ИППИ РАН](http://www.iitp.ru/" \t "https://ruscorpora.ru/page/corpora-about/_blank)], Института лингвистических исследований РАН [[ИЛИ РАН](http://www.iling.spb.ru/" \t "https://ruscorpora.ru/page/corpora-about/_blank)] в  Санкт-Петербурге, [Воронежского государственного университета](http://www.vsu.ru/" \t "https://ruscorpora.ru/page/corpora-about/_blank). Программную и организационную поддержку проекту с его основания оказывает компания «[Яндекс](http://yandex.ru/)». На протяжении истории Корпуса в проекте участвовали сотрудники многих других организаций, независимые исследователи, студенты и аспиранты, волонтёры.Подробная информация о сотрудниках проекта, на разных этапах принимавших участие в работе над Корпусом, размещена на странице «[Участники](https://ruscorpora.ru/page/corpora-team/)».- Чем характеризуется национальный корпус? Каков его состав?Национальный корпус представителен и сбалансирован, т. е. содержит все типы письменных и устных текстов, представленных в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию). В Национальном корпусе русского языка в настоящее время широко используются три типа разметки: метатекстовая (она характеризует текст в целом, с точки зрения его автора, жанровых и других особенностей), морфологическая и семантическая (морфологическая и семантическая разметки характеризуют не целый текст, а отдельное слово).- Зачем нужен корпус языка?Национальный корпус предназначен в первую очередь для научных исследований лексики и грамматики языка. Другая задача корпуса — предоставление всевозможных справок, относящихся к указанным областям (лексика, грамматика, акцентология, история языка). Раньше специалист мог лишь просматривать тексты и вручную выписывать из них нужные примеры; эта деятельность была очень трудоемкой и не позволяла обрабатывать большие массивы материала. Теперь ограничений на объем анализируемого материала и скорость поиска информации нет, а это означает, что в распоряжении ученого и преподавателя оказываются колоссальные массивы текстов самого разного типа. Основными пользователями национальных корпусов являются, конечно, исследователилингвисты самого разного профиля. Однако ими круг пользователей корпуса вовсе не ограничивается! Надежные статистические данные о языке определенной эпохи или определенного автора могут интересовать литературоведов, историков и представителей многих других областей гуманитарного знания. Важнейшее значение национальные корпуса имеют и для преподавания языка.

2. Исследуйте вопрос о частоте использования слова «кофе» в мужском и среднем роде. Проверьте, насколько жизненным является предположение о том, что мужской род слова «кофе» объясняется широким использованием в России в разные периоды двойственного слова «кофий». Используйте временной период - последние 200 лет.

Слово кофий было популярно в 19 веке. Предположение о том, что мужской род слова «кофе» объясняется широким использованием в России в разные периоды двойственного слова «кофий» необоснованно, так как словоформа «кофий» почти не употребляется в последние 200 лет.Пики популярности среднего рода приходятся на конец 19 века, когда большая часть России была занята в промышленном производстве. Полагаю, что сфера грамоты и образования в целом могли серьезно «западать» у русских представителей рабочего и крестьянского классов. Неудивительно, что слово «кофе» было переделано под «лад» обычных русских слов, заканчивающихся на Е и имеющих средний род: солнце, море, поле. Затем в 1930-1940 годы «кофе» среднего рода снова стало популярно. Возможные причины: 1)перестройка орфографии; 2)присоединение республик (Туркменская, Таджикская, Узбекская и т.д.), адаптирование слов под языковые особенности наций. В наши дни словоформа «кофе» среднего рода снова становится популярной.

3. Сравните популярность использования в русском языке слов «компьютер» и «ЭВМ» в последнее столетие. Сделайте вывод о динамике изменений и основных тенденциях.

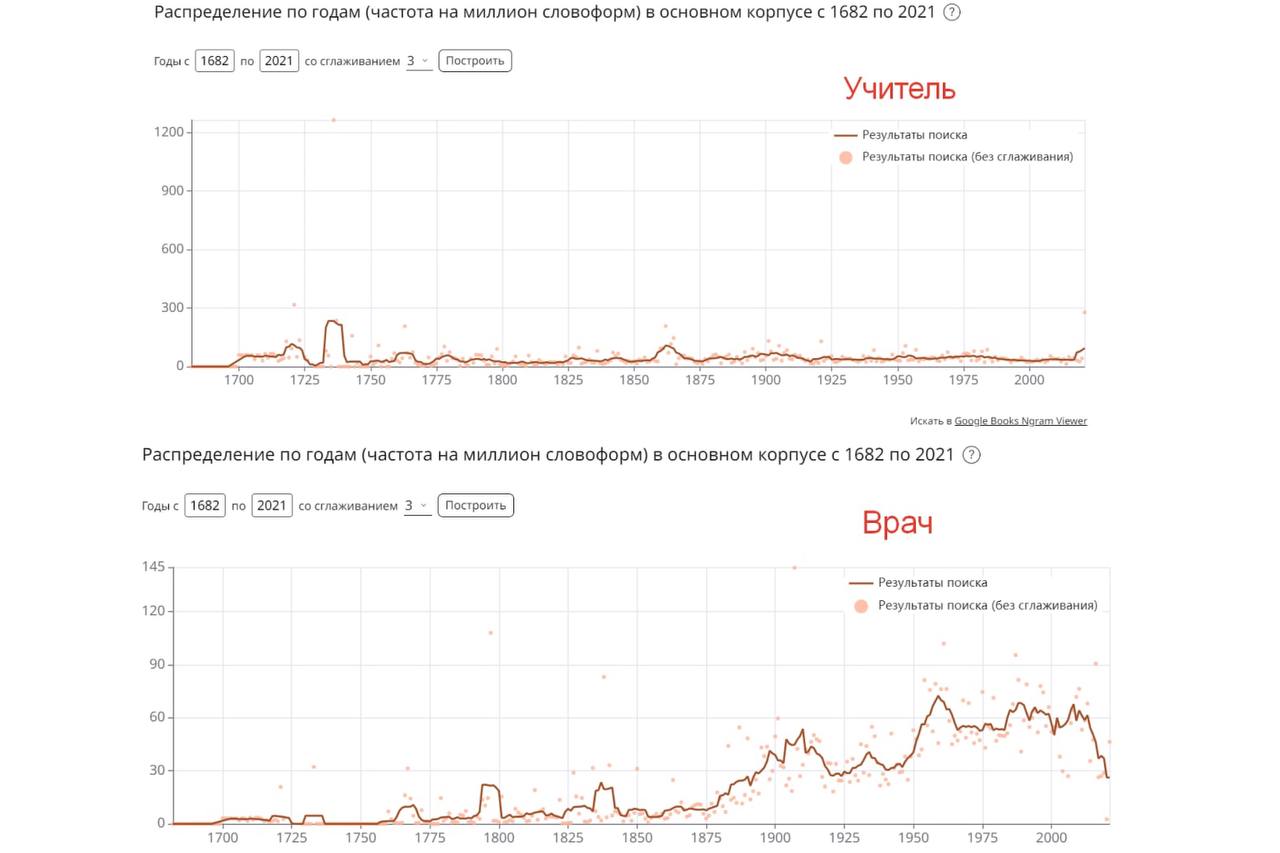
Слова «ЭВМ» и «компьютер» произошли практически в одно время. Это объясняется технической революцией 1970-ых гг., появлением электронно-вычислительных машин. Однако слово «компьютер» буквально вытеснило из обихода «ЭВМ». Лишь с 2022 года последнее снова начинает пользоваться популярностью, предположительно, как сленг среди профессионалов в технической сфере.

ЭВМ



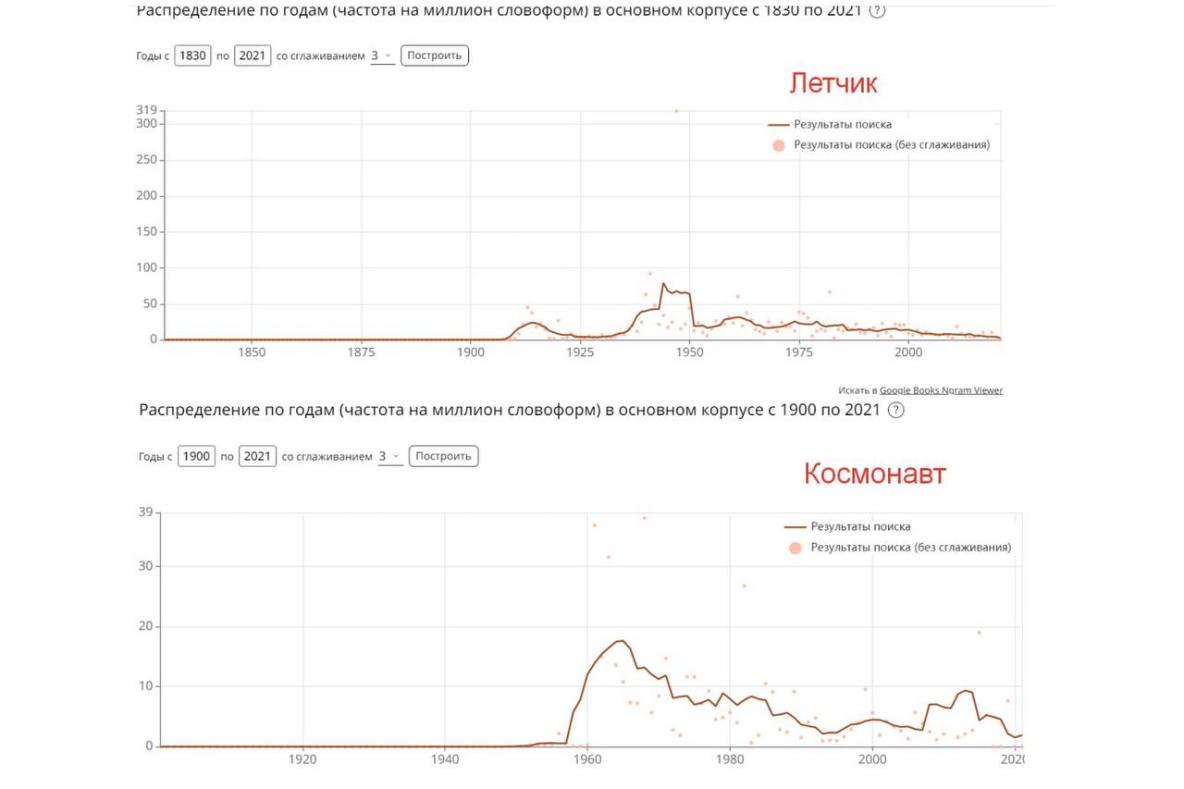
ЭВМ

4. Исследуйте популярность профессий летчик, космонавт, врач, учитель, писатель, фотограф, используя частотный анализ по годам. Постарайтесь найти объяснения «пикам» популярности, связав их с историческими событиями в стране, культурными явлениями или другими причинами. Пики популярности в употреблении слов связаны с актуализацией профессий в конкретный период времени.



1)Пик слова «учитель» приходится на XVIII век, когда в России образование получали дети из семей разных сословий - купеческие имели домашнее образование, дети мещан – в государственных школах и начальных училищах, дети разночинцев обучались в гимназиях и Московском университете. Также появились первые начальные государственные школы и образование получало широкие слои населения.

Пик слова «врач» приходится на времена, когда люди сталкивались с болезнями и эпидемиями. Особенно точно это можно прочувствовать в связи с недавней эпидемией COVID-19, когда профессия врач получила еще большую востребованность, следовательно, участилось использование этого слова.



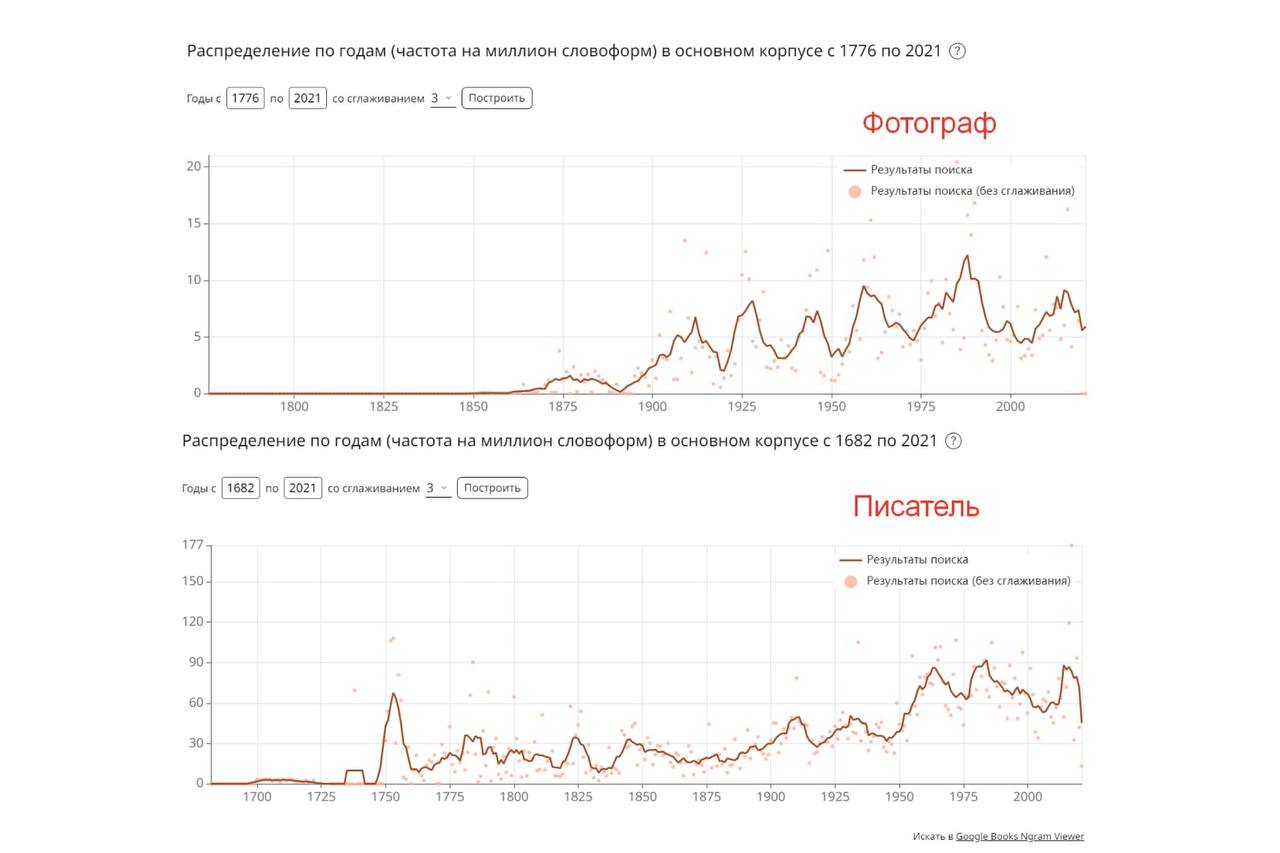
2)Появление слова «космонавт» накладывается на пики в его использовании и приходятся на 1960-ые годы:4 октября 1957 года – начало космической эры – был запущен первый искусственный спутник Земли (ПС-1).

12 апреля 1961 года советский космонавт Юрий Гагарин на космическом корабле «Восток-1» стартовал с космодрома «Байконур» и впервые в мире совершил орбитальный облёт планеты Земля. В 60-ю годовщину исторического события депутаты Мосгордумы поздравили москвичей и всех жителей страны с Днем космонавтики.

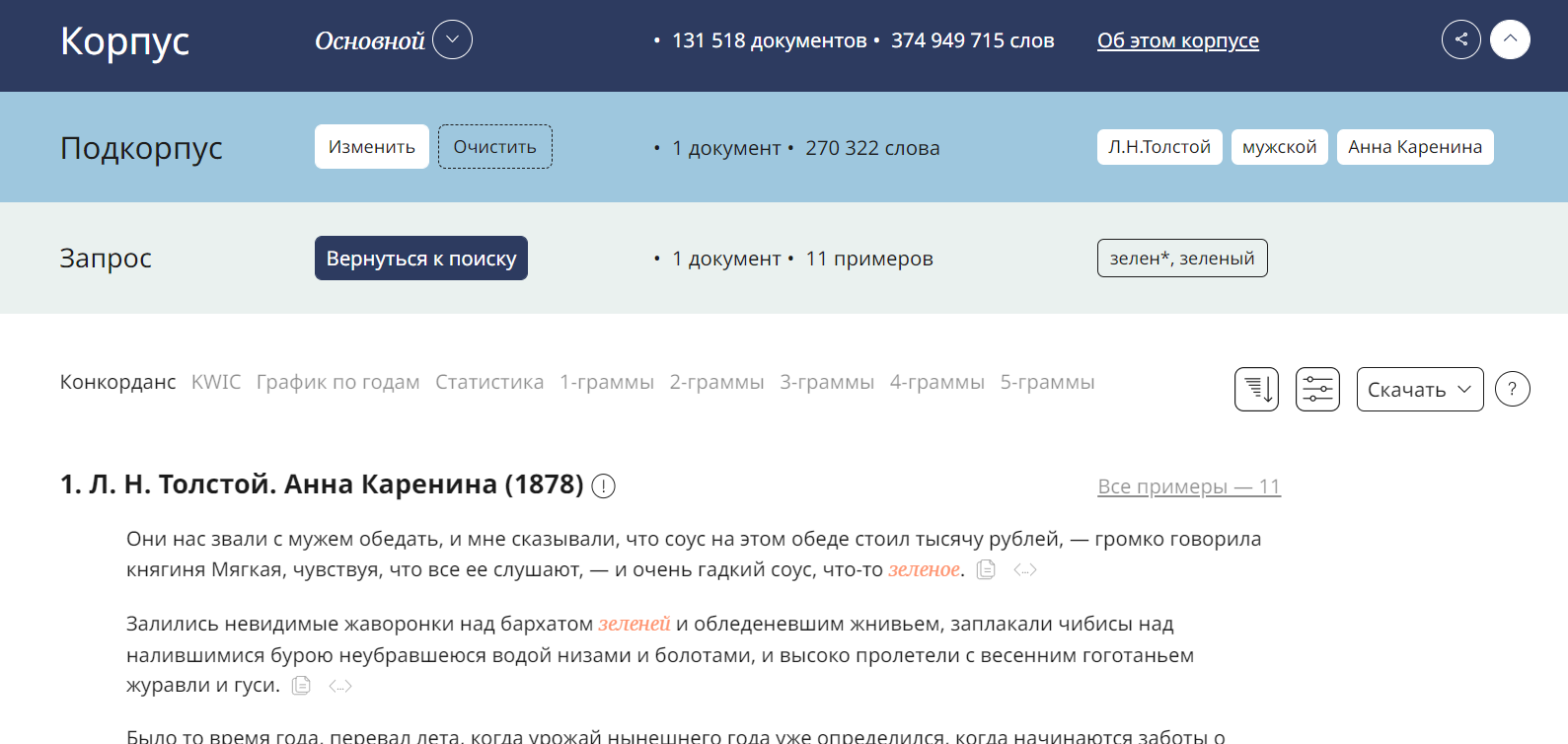
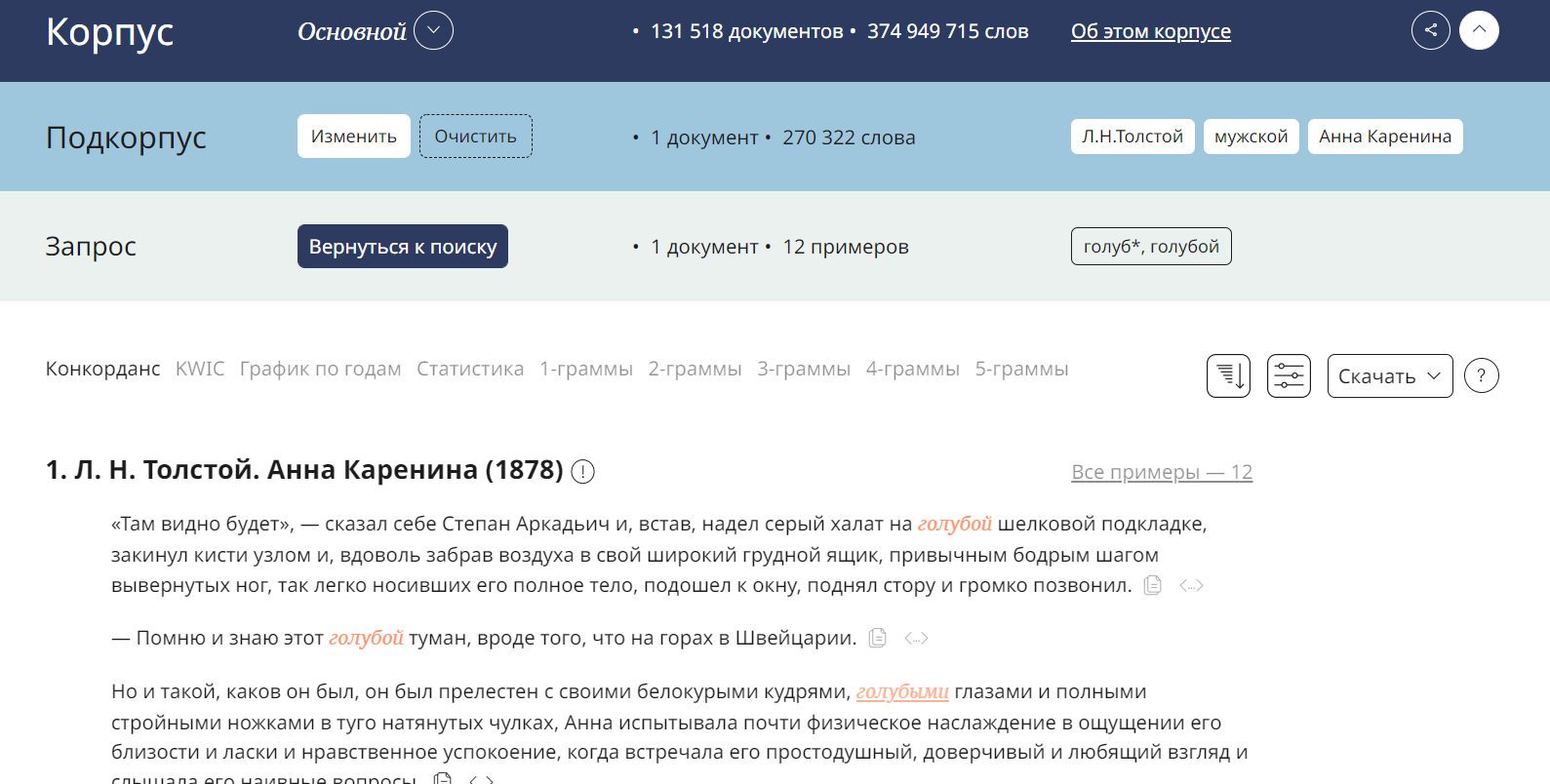
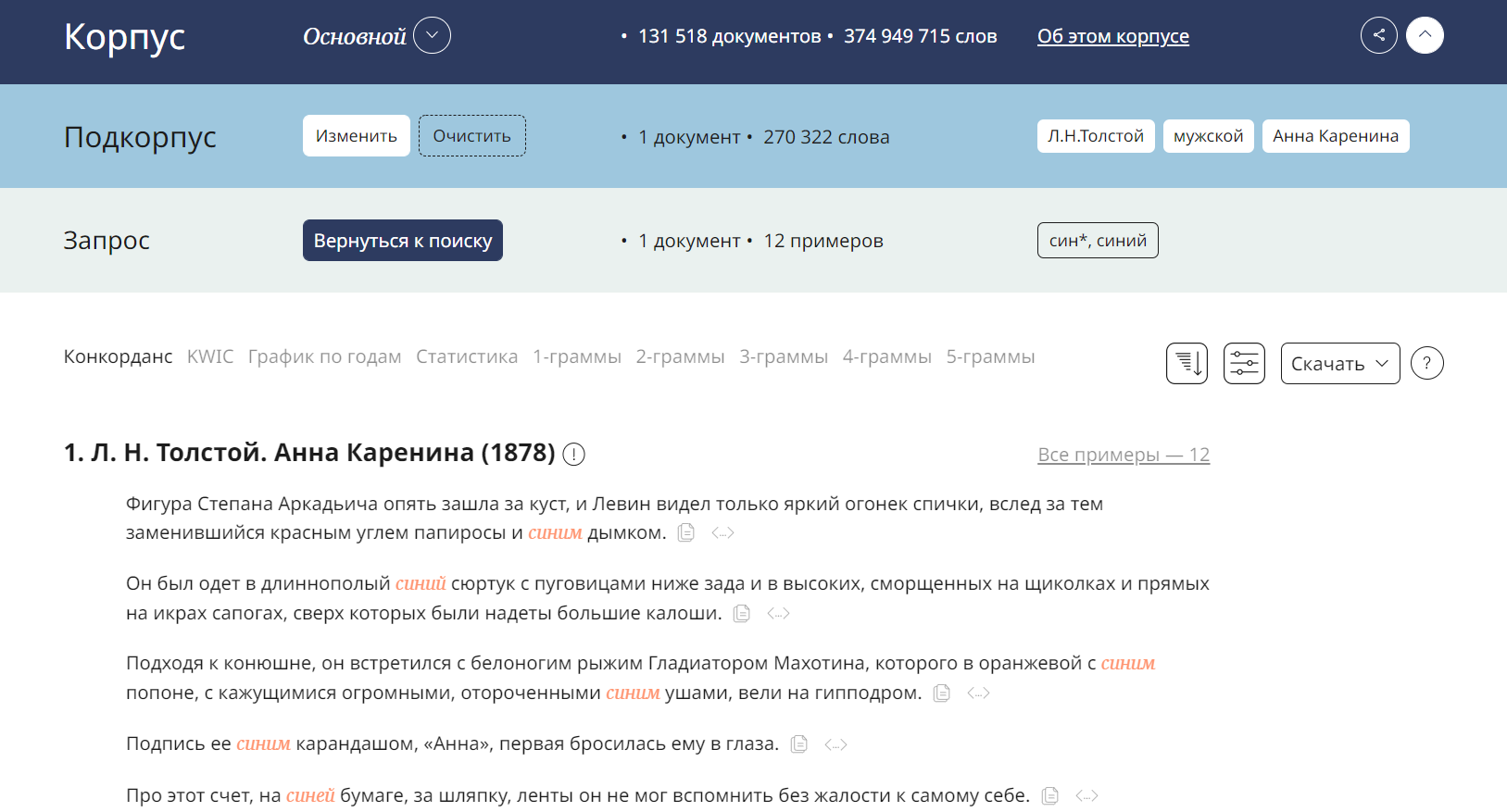
Пик употребления слова «летчик» приходится на 1940-ые годы. Это время, когда большая часть стран и народов Европы, Азии принимали участие во Второй мировой войне. В этот «кровавый» период истории был всплеск военных профессий. Летчик является одной из таких.

3)Наибольшую популярность сейчас получают такие творческие профессии, как фотограф и писатель. Именно поэтому пики их употребления приходятся на последние 100-200 лет. Интересно, что «фотограф» было частотно в 1970-80ые гг., что связано с технической революцией, появлением новых камер и фотоаппаратов.

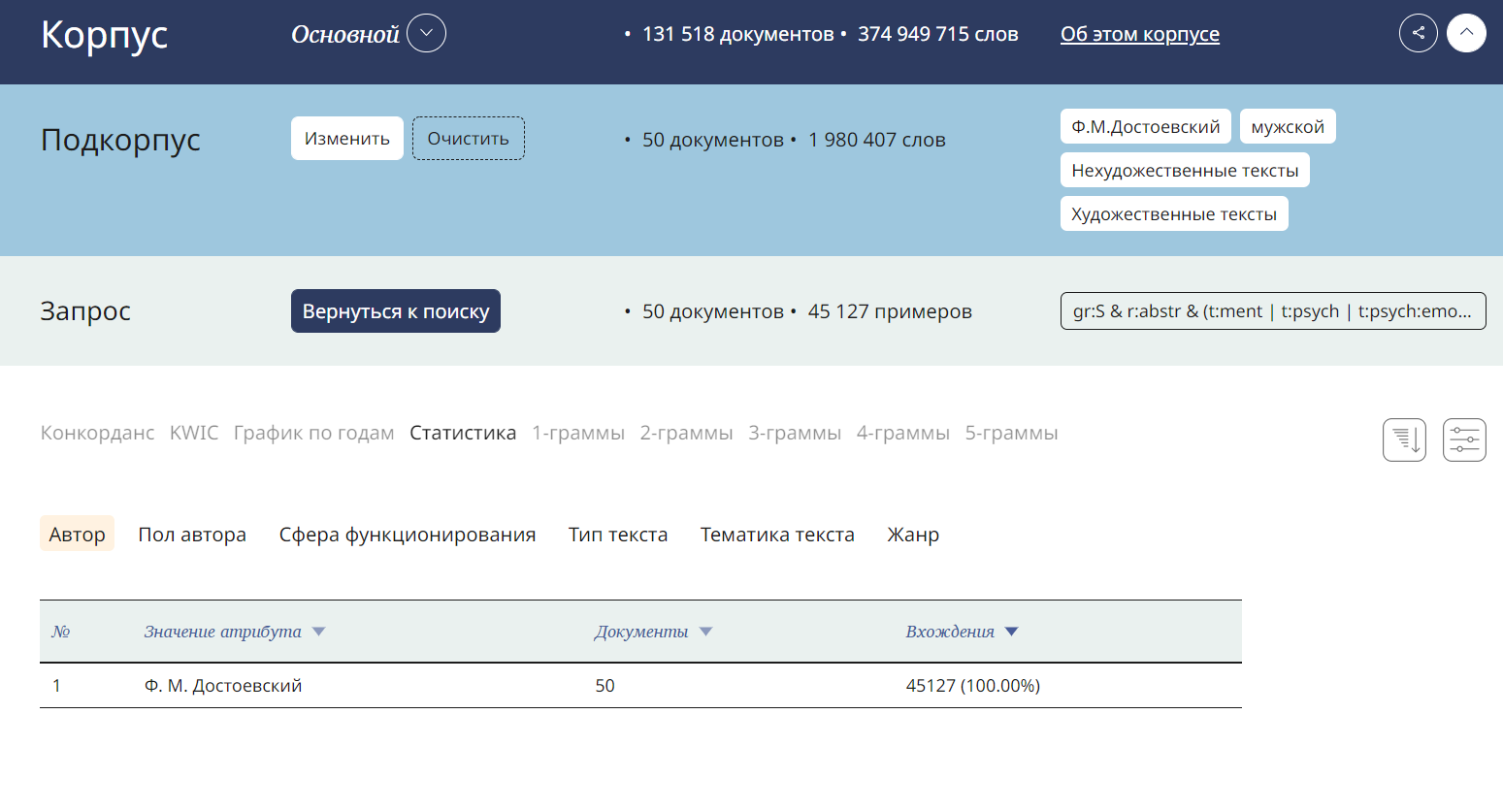
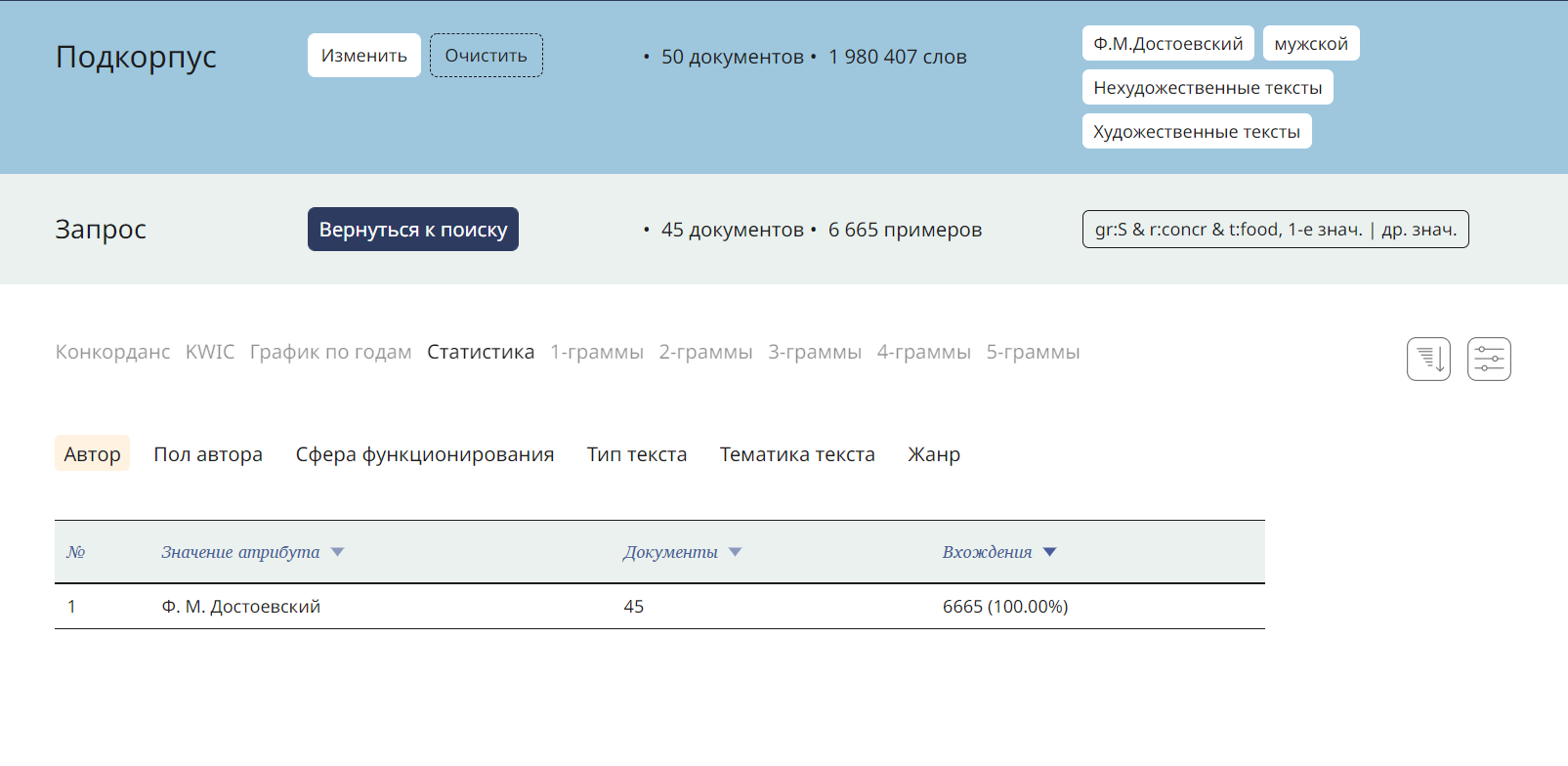
А слово «писатель» отмечено яркой вспышкой употребления в 1750-ые гг., что также переплетается с историей. 1741 - 1761 - Царствование Елизаветы Петровны, в годы правления которой: 1)1750 - Открытие в Ярославле первого русского театра (Ф.Г. Волкова) 2)1755 - Основание Московского университета 3)1757 - Учреждении Академии художеств. Всё это несомненно способствовало росту духовной составляющей человека, появлению людей, желающих выразить свои эмоции через слово.



5.Проверьте, насколько оправдано решение об использовании преимущественно сине-голубых оттенков при экранизации романа Л.Н.Толстого «Анна Каренина» британским режиссером Джо Райтом (экранизация 2012 года с Кирой Найтли в главной роли). Связано ли выделение синего цвета, как доминирующего, в кадрах этого фильма с оригиналом текста романа. Для этого проверьте частоту использования слов «синий», «голубой» в романе Л.Н.Толстого по сравнению со словами, обозначающими другие цвета.

По окончании исследования, мы видим, что синий употреблено 12 раз,голубой - 12 раз, красный - 58, зеленый - 11 раз, желтый - 16 раз. Таким образом, выделение синего цвета, как доминирующего в кадрах фильма, не связано с оригиналом текста романа. 

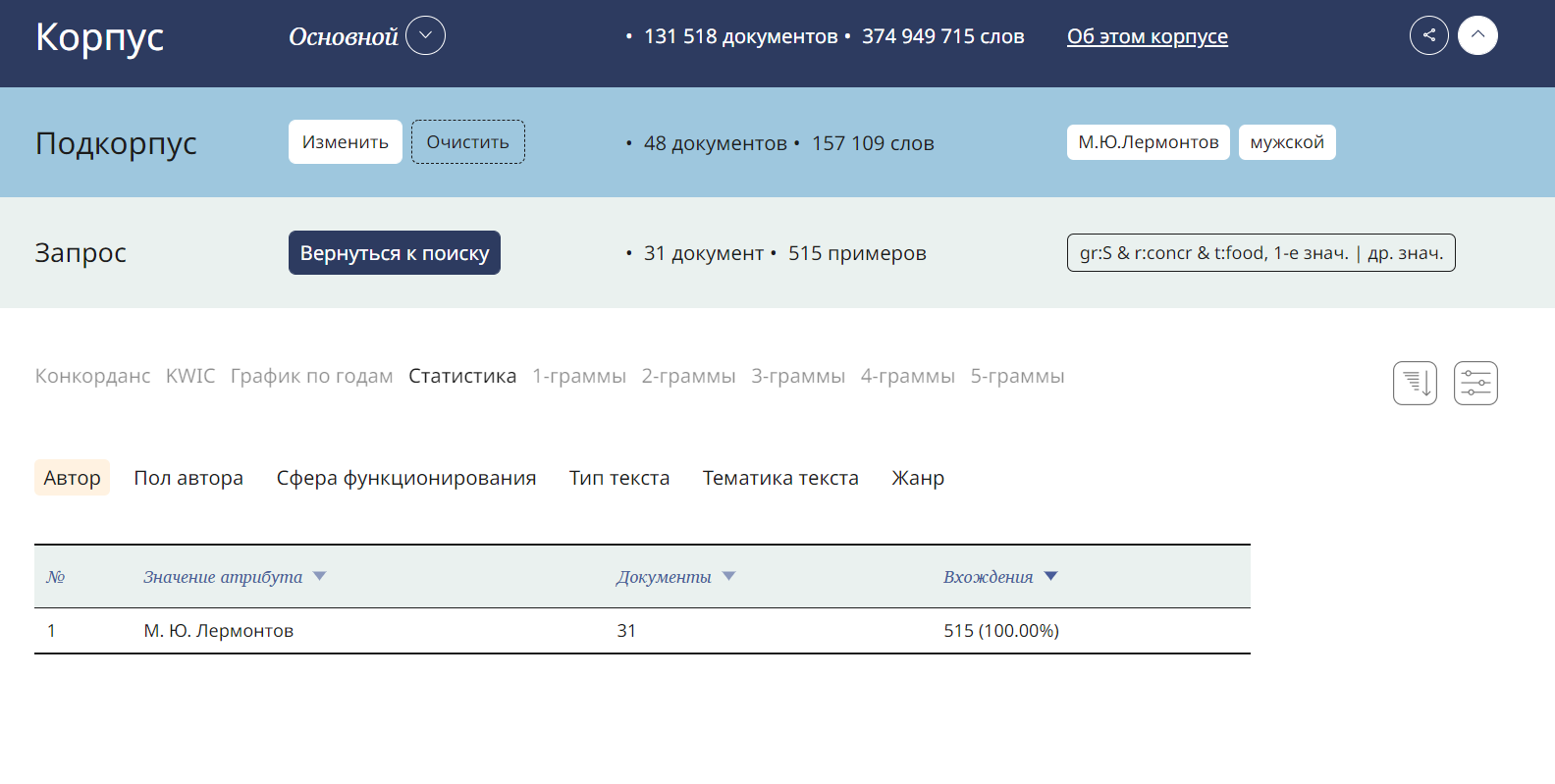
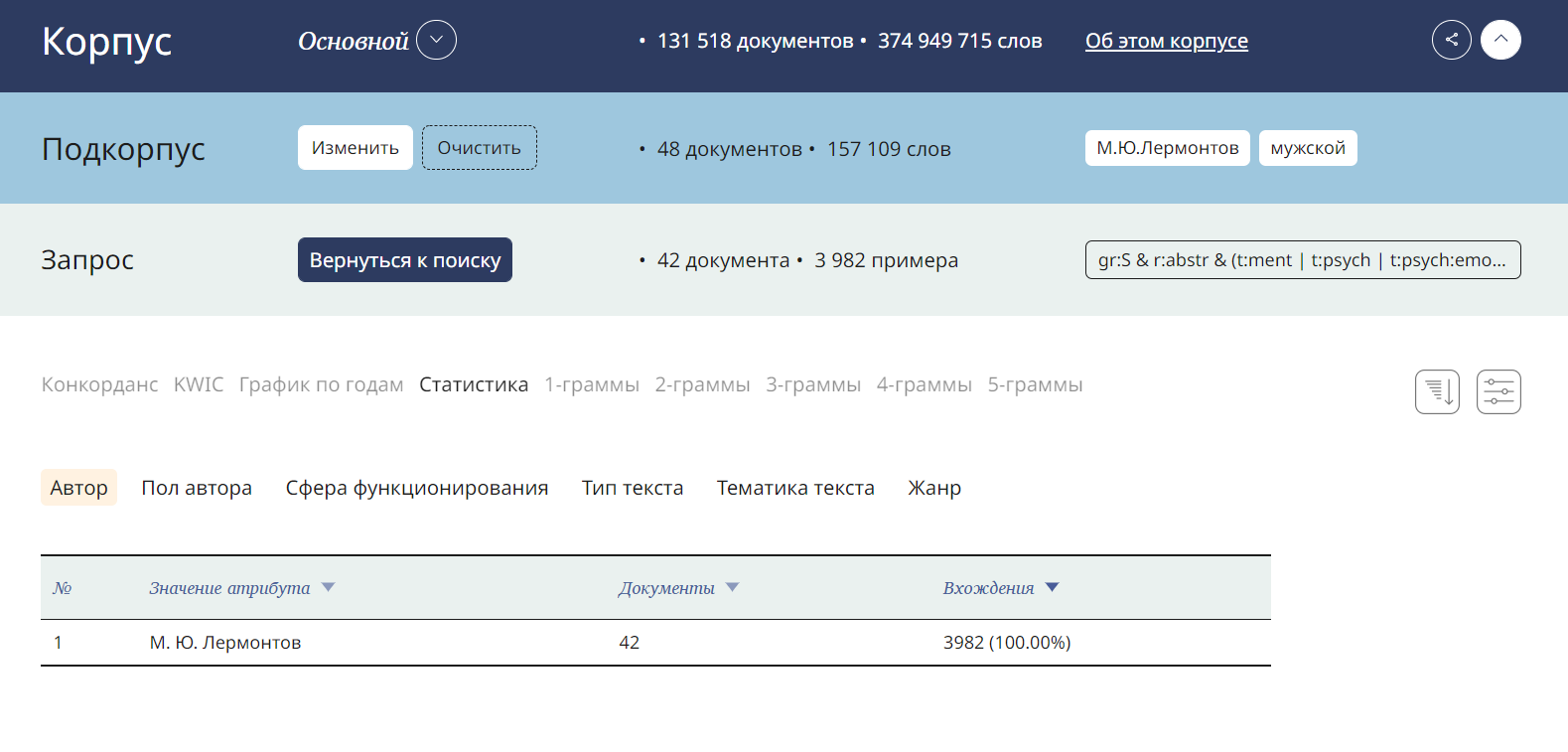
6.Проверьте, насколько оправдано представление о Ф.М.Достоевского, как писателе, пишущем о бестелесном, духовном, психологическом, мало говорящем о бытовом и земном. Для этого выделите подкорпус с его произведениями и проведите частотный анализ семантических признаков, выбрав полярные значения «еда и напитки» (предметные признаки) и «психическая сфера» (непредметные признаки). Аналогичное исследование проведите для других известных писателей (хотя бы с одним), сравните полученные результаты.

 Количество примеров в категории «еда и напитки»: 6665

Количество примеров в категории «ментальная сфера» и «психическая сфера»: 45127

Предположение о том, что Ф.М.Достоевский писатель, пишущий о бестелесном, духовном, психологическом, мало говорящем о бытовом и земном, подтверждается.

Исследование для М.Ю.Лермонтова:



Количество примеров в категории «еда и напитки»: 515

Количество примеров в категории «ментальная сфера» и «психическая сфера»: 3982

Несмотря на то, что у М.Ю.Лермонтова категории, связанные с психологическим состоянием человека превалируют над бытовыми вещами, у Ф.М.Достоевского намного больше слов, выражающих духовную составляющую личности.