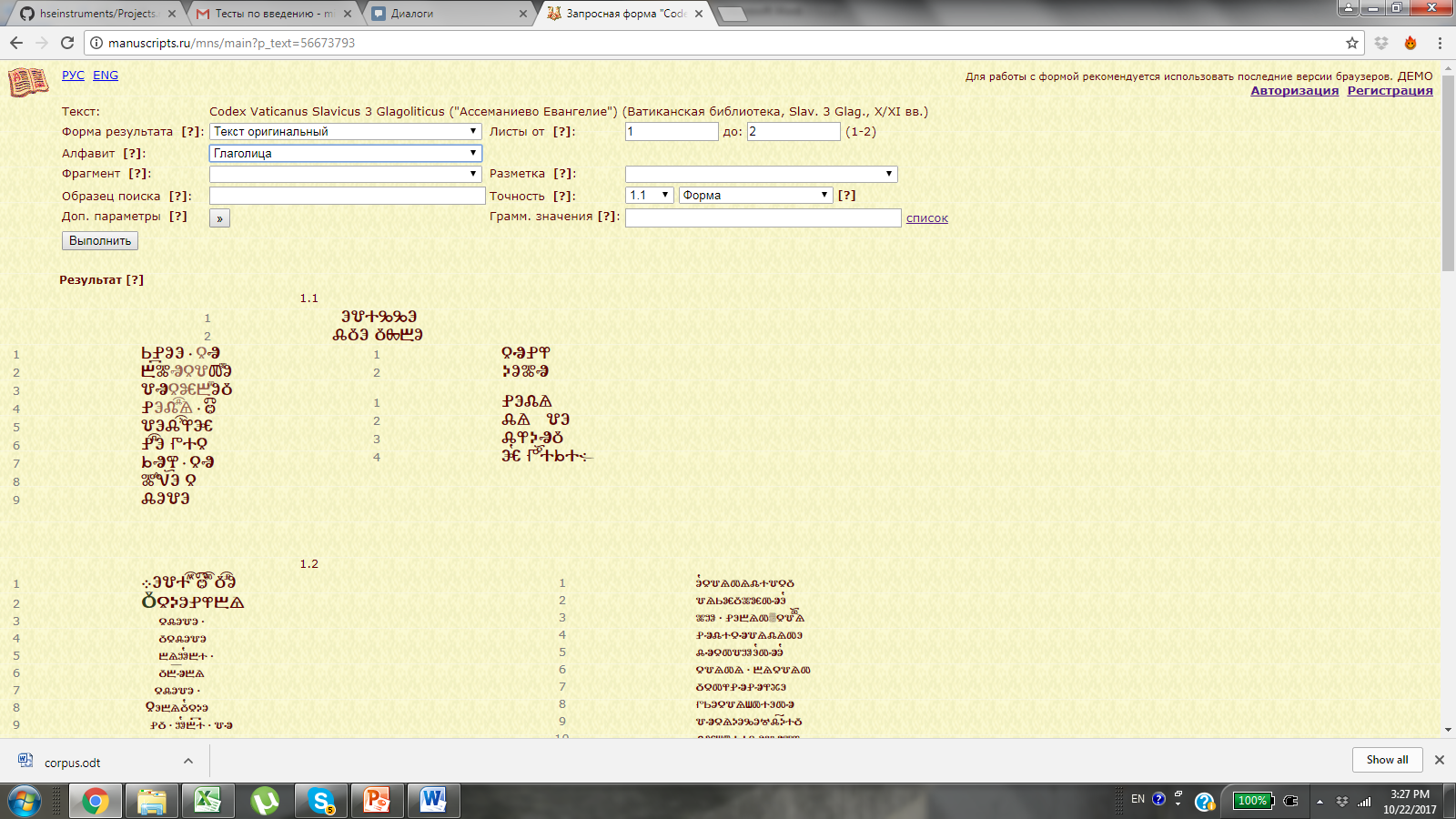
**Работа с корпусом славянского письменного наследия «Манускриптъ» (manuscripts.ru). Отчет о возможностях, достоинствах и недостатках**

Михаил Браверман ([mika.uchenik@gmail.com](mailto:mika.uchenik@gmail.com)), Фаина Даниэль ([faina.daniel@gmail.com](mailto:faina.daniel@gmail.com))

Сайт manuscripts.ru содержит 18 независимых корпусов – в основном старославянских и древнерусских текстов, а также корпус поэзии М.В.Ломоносова и корпус удмуртской поэзии (корпус создан Удмуртским государственным университетом, видимо, это области, в которых работают лингвисты оттуда). Корпус Ломоносова недоступен незарегистрированным пользователям, а в остальных для просмотра и поиска доступны только первые два листа рукописи (мы даже зарегистрировались и запросили доступ ко всему, но с тех ничего не произошло). Тем не менее, все функции корпусов в демо-версии представлены.

Найти корпус было несложно (на него есть ссылка в НКРЯ, большое количество отчетов о его создании и работе в Google Scholar), а вот разобраться в его работе получилось совсем не сразу. Стилически, дизайн как раз очень высокого уровня и вполне приятный (желтовато-бежевая цветовая гамма, например, красиво отражает то, что корпус посвящен древним рукописям). А вот в содержательном плане интерфейс довольно бардачный, особенно для новичка. Например, на основной странице сайта есть ссылки на корпуса сайта, но почему-то не на все, тогда как на другой странице есть абсолютно тот же список, только полный. В поиске тоже без поллитра не разберешься (хотя, наверное, он вызвал бы меньше проблем у специалистов по старославянскому/древнерусскому, знакомых со стандартными формами передачи рукописных текстов на экране и т. п.). Подсказки есть, но зачастую написаные так, что помощи от них мало. Это все, однако, не отменяет того, что в корпусе есть множество очень полезных функций, продуманных с учетом специфики материала корпуса. Конкретные примеры привожу ниже, по ходу рассмотрения устройства поиска.

Итак, рассмотрим устройство поиска на примере корпуса глаголических рукописей X-XI вв. (в остальных корпусах все устроено примерно так же, с некоторыми дополнительными особенностями, например, в корпусе древнерусских переводов с греческого есть параллельный корпус, и т.д.). Открываем корпус, перед нами две возможности: «В многотекстовую запросную форму» или «В запросную форму текстов». Нажимаем на второе, открывается список текстов внутри данного корпуса, выбирая один из них, мы открываем старницу поиска внутри данного текста (в нашем случае – внутри первых двух листов текста). Возьмем для примера поиск внутри «Ассеманиевого Евангелия».

Как можно видеть, вид текста максимально приближен к виду самой рукописи, учтено распределение по столбцам, разный размер букв в разных в местах, лигатуры, диакритики и даже различно отображаются буквы, присутствующие в тексте в нескольких вариациях (ср. фотографию первого листа рукописи upload.wikimedia.org/wikipedia/commons/f/f9/CodexVaticanusSlavicus3Gagoliticus.jpg). Серым отмечены буквы, которые тяжело разобрать в изначальном тексте. Для удобства работы с текстом, в графе «Алфавит» можно выбрать транскрипцию текста на старославянской кириллице (причем в 5 разновидностях, переводящих лигатуры и диакритики в аналогичные кириллические!), современной кириллице или латинице. Кроме того, в графе «Форма результата» можно выбрать значение «Текст преобразованный», которое вставляет словоразделительные пробелы и упрощает некоторые другие графические сложности изначального текста. Эта же графа отвечает, собственно, за форму результата (подчеркивание в тексте или вывод отдельно словоформы, например).

Обратимся теперь к устройству самого поиска. Вне зависимости от выбранного алфавита, поиск производится на старославянской кириллице, клавиатура которой выводится на экране при вводе в этом поле (в смысле таблица с символами, на которые нужно кликать). Также при поиске в таблице в поле «Точность» можно выбрать разную точность: во-первых, можно искать по форме или лемме, а во-вторых, существует отдельная разметка, обознчаемая числами, связанная с графическим разнообразием старославянского написания (там, где в скриншоте выбрано 1.1). Здесь при нажатии на знак вопроса рядом с графой вылезает как раз довольно полезная и self-explanatory подсказка:

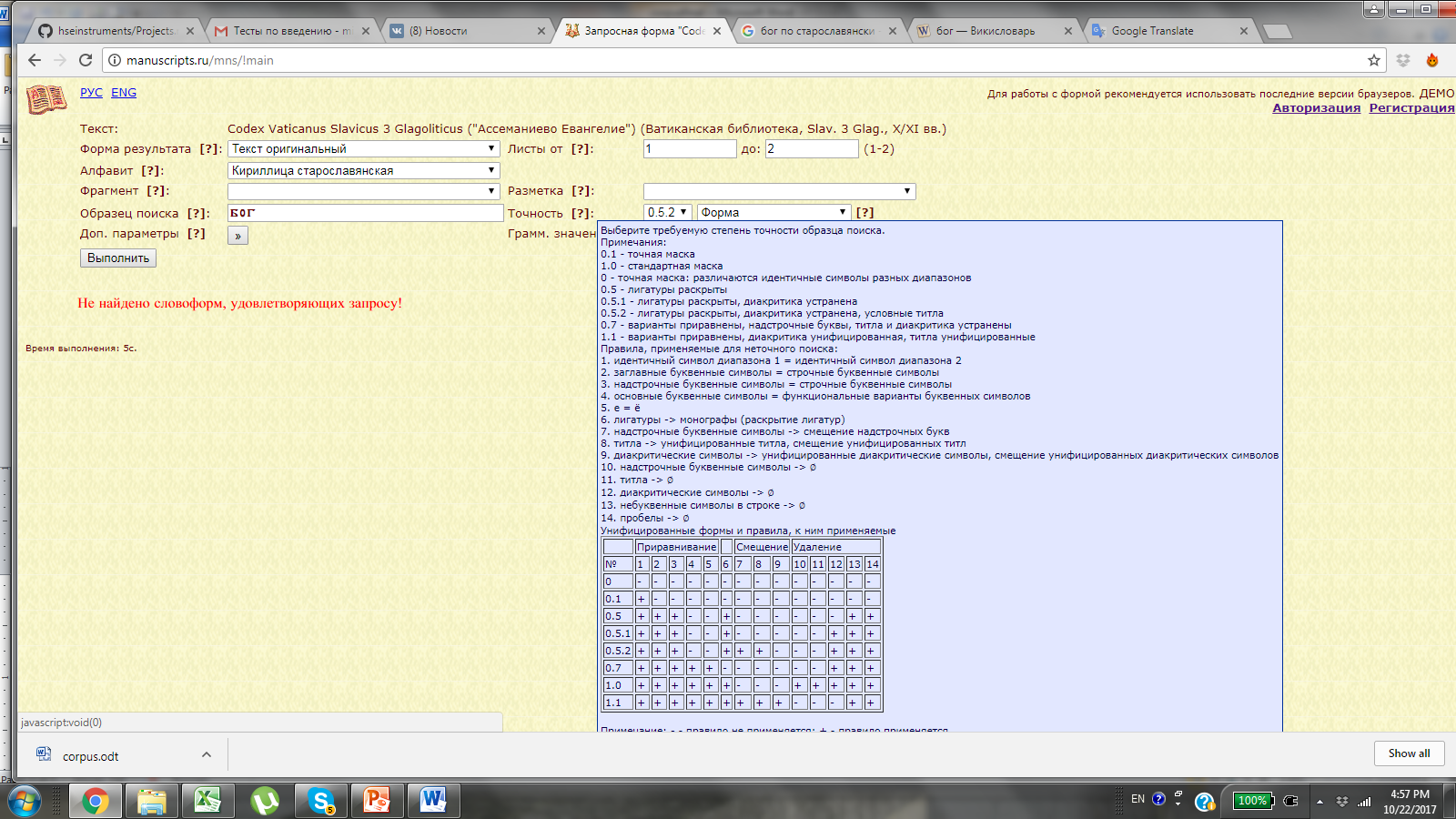
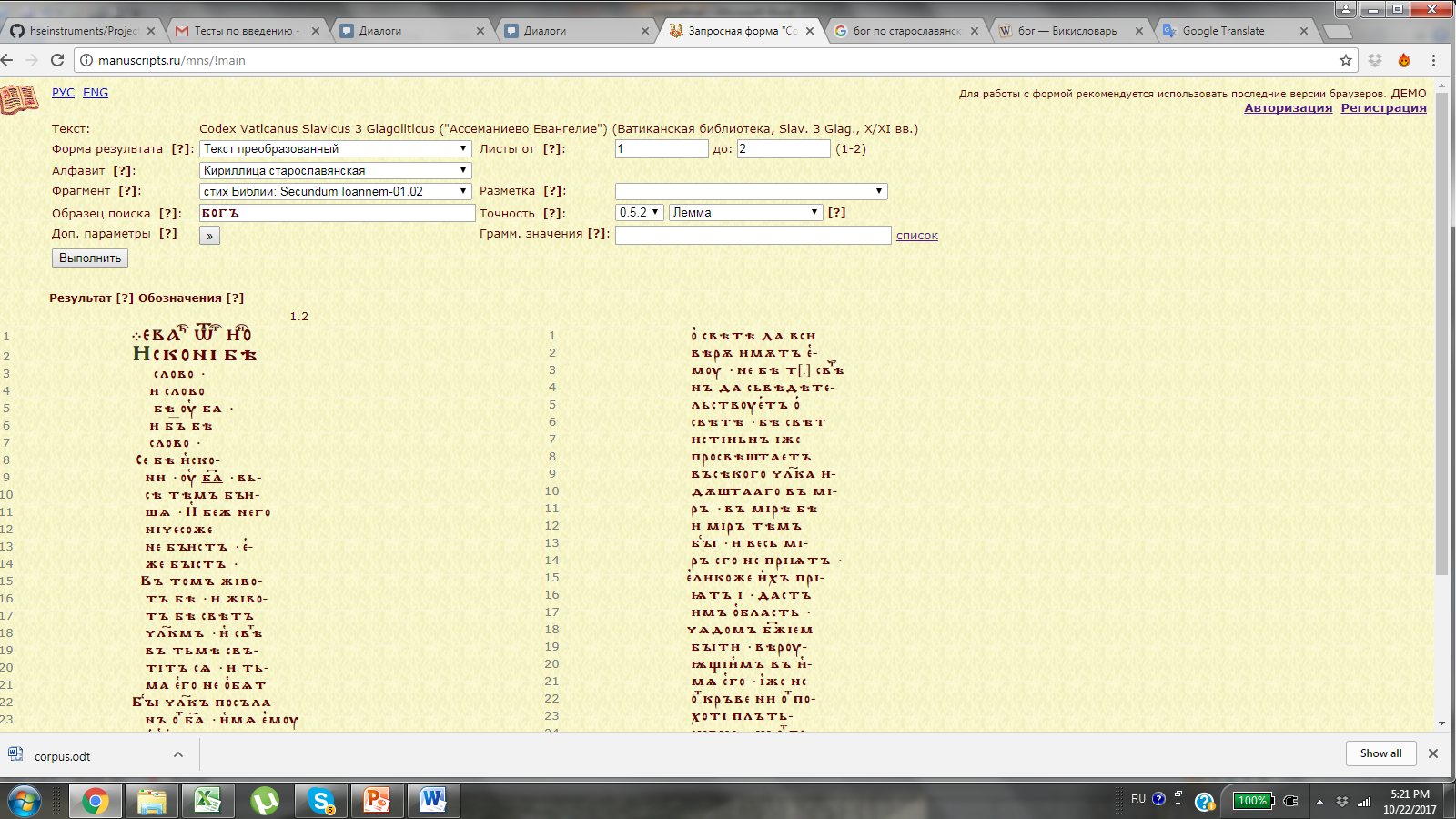


Таблица объясняет, при каких значениях точности какие параметры учитываются, а какие – нет. Эта безусловно весьма адекватная и удобно сделанная разметка для корпуса древних текстов. Сами численные значения, правда, выглядят взятыми с потолка (особенно мне нравится, что, вроде как, значения – числа от 0 до 1, при этом есть значение 1.1), но это уже придирка (функционально же никак не сказывается).

Графа «Доп. параметры» состоит из трех параметров: «Тип текста», «Модификации» и «Дефекты». Первый позволяет ограничить область поиска основным текстом, или маргиналиями, или таблицами, etc.; второй – выбрать разные графические параметры искомого (обычный шрифт, полужирный, верхняя перекладина, etc.); третий – учесть различные возможные графические дефекты. Смысл остальных полей очевиден («Грамм. значения» представляют собой упрощенную грамматическую разметку НКРЯ).

**Пример поиска:**



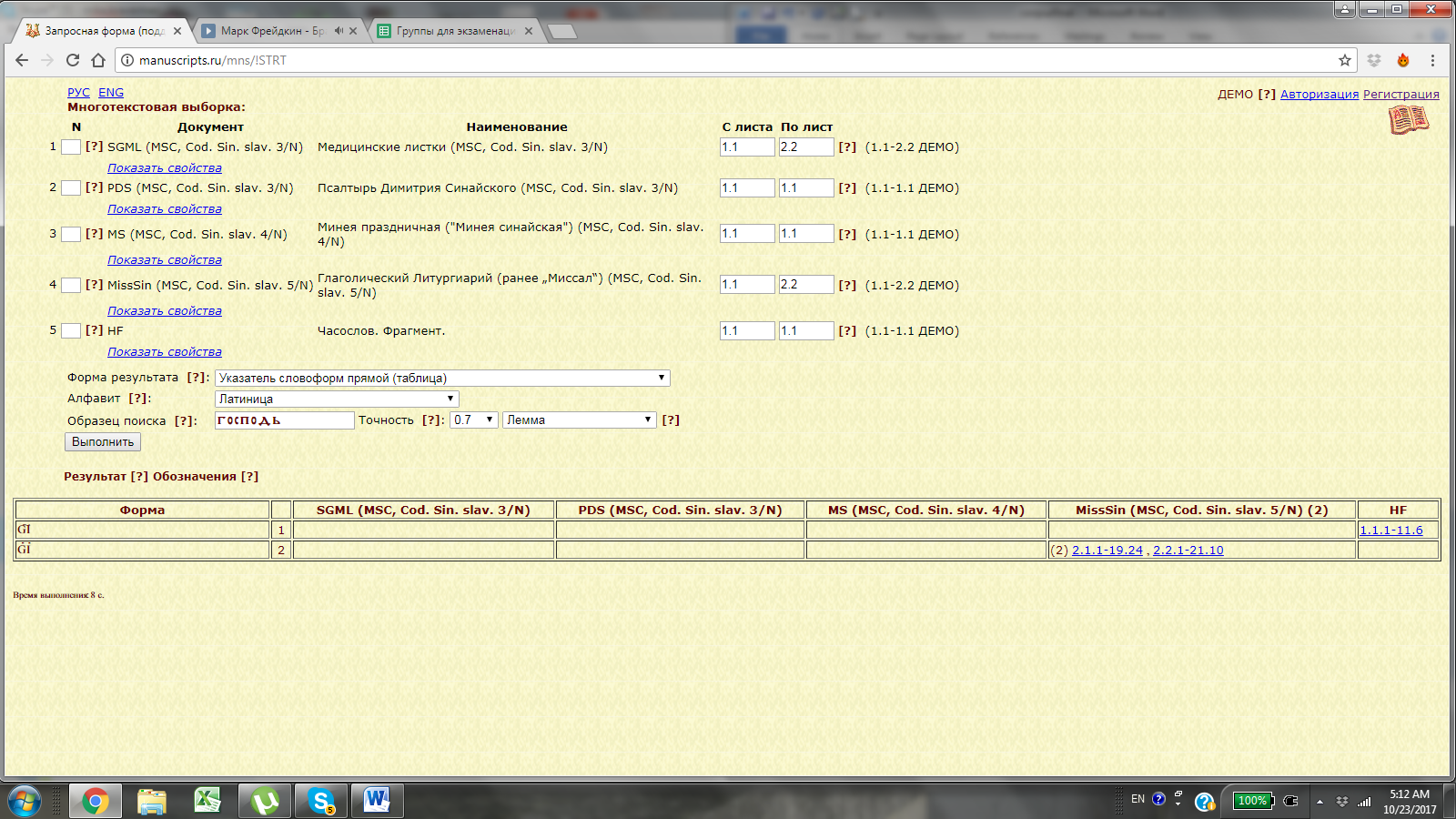
Поиск леммы «богъ» в преобразованном тексте в стихе 1.2 в старославянской кириллице с точностью 0.5.2. Стих 1.1 – какое-то авторское предисловие, а стих 1.2 – уже начало евангелия от Иоанна, то есть «В начале было Слово, и Слово было у Бога, и Слово было Бог. Оно было в начале у Бога...». Легко видеть, что в старославянском тексте первое «Бога» записано как «ба», «Бог» - как «бъ», а второе «Бога» - как «ба» под титлом. Однако, поиск находит и подчеркивает только последнее, так как точность 0.5.2 раскрывает титла, но не сокращения, записаные просто так. Собственно точности, которая бы раскрывала просто авторские сокращения мы вообще в таблице не нашли – и это явный недочет корпуса (хотя, возможно, это сделано специально, из соображений, что в случае титла мы точно знаем, что это сокращение, а расшифровка авторских сокращений – это уже как бы часть анализа, тогда как корпус – просто база данных). Есть, кстати, еще одна удобная возможность при наборе искомой леммы/формы, которая, казалось бы, должна спасать, но и она не помогает: при наборе можно ввести знак «\_» или «%» – это «джокеры», первый для одной буквы, второй – для любого количества (удобно для поиска по морфемам или по словам, имеющим вариативное написание).

Обратимся теперь к поиску в «многотекстовой запросной форме». Сначала вообще ничего непонятно, потому что перед нами единственная строка «Запрос».

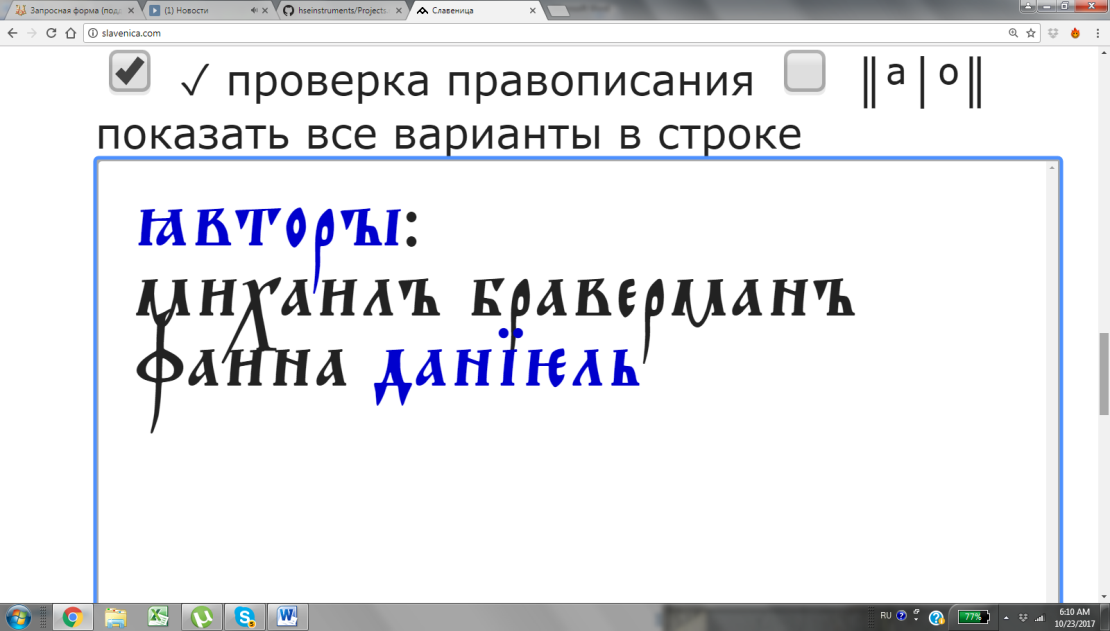


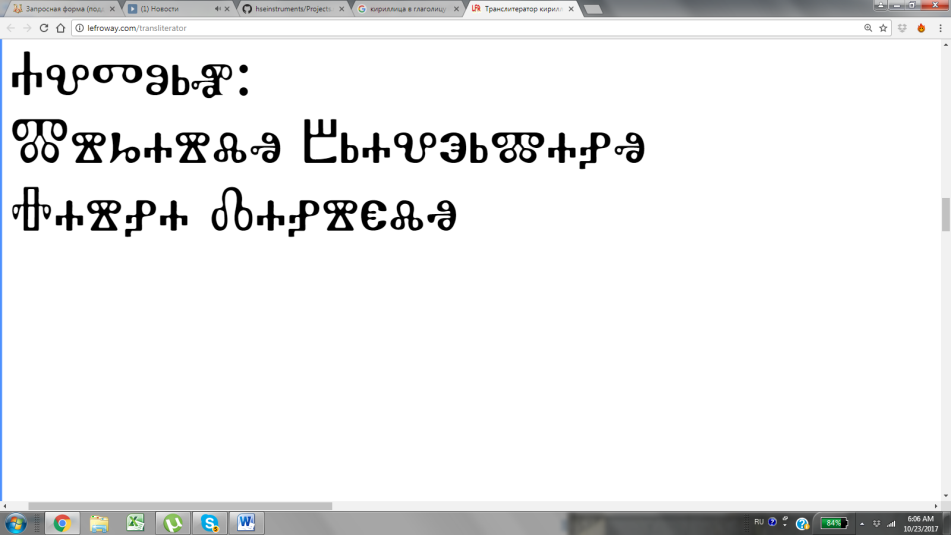
*Для любителей минимализма в веб-дизайне! (это мы сюда вставили, если что)*

Исходя из стандартного устройства корпусов, логично предположить, что это поиск (то ли формы, то ли леммы) по всем текстам корпуса. Однако оказывается, что это поиск по названию текста, а дальше уже есть опция «Искать в найденом», а для того чтобы искать сразу во всех текстах корпуса, в «Запрос» нужно вбить знак %, а потом уже искать в найденном. И это все вроде как объяснено в подсказке [?], но так, что разбирались мы минут 40. При переходе в поиск по выбранным текстам все устроено так же, как при внутритекстовом поиске, но почему-то многие полезные функции отсутствуют.

**Пример поиска:** 

Здесь мы запросили тексты, содержащие в названии или свойствах «синайский». Далее мы по выбранному подкорпусу пробили лемму «господь» с выводом на латинице в форме таблицы (оба раза в одном тексте, в форме «господи» под титлом). Как видно, многих параметров, которые были во внутритекстовом поиске, отсутствуют здесь, и главное совершенно непонятно, почему.

 Итак, мы рассмотрели основные возможности корпуса. Поначалу корпус казался нам очень неудобным (в первую очередь с точки зрения интерфейса и формулировок подсказок), однако разобравшись, он представляется нам наоборот весьма качественным и профессионально сделанным. Действительно, в нем есть некоторое количество слабо продуманных мест, но с другой стороны, раз мы разобрались за сутки и смогли сделать такой высококачественный отчет (простите, самолюбие – оно такое, ничего с ним не сделаешь[[1]](#footnote-1)), это и не так плохо. Кроме упомянутого, для новичка неудобно отсутствие в корпусе переводов текстов на современный русский, хотя, опять же, корпус рассчитан на специалистов, и им, наверное, было бы проще сориентироваться даже при первом заходе на корпус. И все таки в корпусе есть кучу всего, продуманного специально для данной области исследований. Короче говоря, в итоге впечатление положительное.



1. Кстати, у нас имеется много несданных в срок ДЗ, и отчетом мы старались максимально компенсировать, и нам правда кажется, что он заслуживает того, чтобы формула оценки была подкорректирована в пользу учета экзамена с большим коэффициентом. Извините за наглость, но обещаем работой в следующих модулях доказать, что Ваша благосклонность оправдана и в этом модуле был случайный сбой с непривычки к университету. [↑](#footnote-ref-1)