Русская речь
Russian Speech

№ 05 | 2024
С. / Pp. 34–47

Из истории русского языка

Текстологические исследования эпохи big data и нейронных сетей

Александр Геннадьевич Кравецкий[1], Светлана Михайловна Кусмауль[2], Екатерина Андреевна Мишина[3], Александра Андреевна Плетнева[4],
Институт русского языка им. В. В. Виноградова РАН (ИРЯ РАН) / Национальный исследовательский ядерный университет «МИФИ» (НЯУ МИФИ) (Россия, Москва), krav62@mail.ru[1], kusmauls@yandex.ru[2], kmishina@mail.ru[3], apletneva@list.ru[4]

DOI: 10.31857/S0131611724050035

аннотация: В статье анализируются возможности, открывающиеся перед филологами-древниками благодаря появлению технологий работы с большими данными. Речь идет об изучении и издании текста древнерусских рукописей традиционного содержания, необходимых для совершения богослужения. Эти рукописи существовали в огромном количестве списков и в процессе переписывания подвергались значительной текстовой унификации, что крайне затрудняет их изучение методами традиционной текстологии, опирающейся на трудоемкий анализ разночтений. Сейчас, когда появилась возможность автоматической обработки полного текста памятников, началась работа над созданием системы «Лингвистическая интеллектуальная среда» (ЛИС) — инструмента, который предоставит ряд принципиально новых возможностей для исследования славянских богослужебных текстов разных эпох. В результате будет создан корпус богослужебных текстов XI–XVII вв., полученных с помощью программы по автоматическому распознаванию текста рукописей, с разметкой и поиском. Для каждого фрагмента богослужебной книги пользователь ЛИС будет иметь возможность получить полный перечень разночтений по максимально широкому кругу рукописей. Фактически речь идет о новом типе издания памятников традиционного содержания с возможностью задавать параметры этого издания в соответствии со своими исследовательскими интересами.

КЛЮЧЕВЫЕ СЛОВА: история славянских языков, церковнославянский язык, текстология, богослужебные книги, цифровые методы в гуманитарных науках

ДЛЯ ЦИТИРОВАНИЯ: Кравецкий А. Г., Кусмауль С. М., Мишина Е. А., Плетнева А. А. Текстологические исследования эпохи big data и нейронных сетей // Русская речь. 2024. № 5. С. 34–47. DOI: 10.31857/S0131611724050035.

БЛАГОДАРНОСТИ: Работа выполнена в рамках программы Приоритет 2030 НИЯУ МИФИ.

From the History of the Russian Language

Textual Studies of the Era of Big Data and Neural Networks

Alexandr G. Kravetskiy[1], Svetlana M. Kusmaul'[2], Ekaterina A. Mishina[3], Alexandra A. Pletneva[4], V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences / National Research Nuclear University "MEPhI" (Russia, Moscow), krav62@mail.ru[1], kusmauls@yandex.ru[2], kmishina@mail.ru[3], apletneva@list.ru[4]

ABSTRACT: This article analyses the emergence of new technologies for working with big data which can be highly helpful to philologists, studying the diachronic development. First of all, that applies to the study and publication of Old Russian manuscripts with traditional liturgical texts used for church service. These manuscripts existed in a huge number of folios, and in the process of copying were subjected to considerable textual unification. That makes it very difficult to study them by the laborious methods of traditional textual criticism. Now, when the full text of the monuments can be automatically processed, the creation of the Linguistic intellectual environment (LIE) has been lounged. This tool will provide new opportunities for the study of Slavonic liturgical texts from different historical periods. As a result, we will create a corpus of liturgical texts of the 11th–17th centuries, obtained using a program for automatic text recognition of manuscripts, with annotation und search module. The user of the LIE will be able to receive

a complete list of variant readings for each fragment of a liturgical book of the widest range of manuscripts. In fact, we are talking of a new type of publication of traditional liturgical texts, when the user can set the parameters for the edition in accordance with his research interests.

KEYWORDS: history of Slavic languages, Church Slavonic language, textual criticism, liturgical books, digital methods in the humanities

FOR CITATION: Kravetskiy A. G., Kusmaul' S. M., Mishina E. A., Pletneva A. A. Textual Studies of the Era of Big Data and Neural Networks. Russian Speech = Russkaya Rech'. 2024. No. 5. Pp. 34–47. DOI: 10.31857/S0131611724050035.

ACKNOWLEDGMENTS: This work was carried out within the National Research Nuclear University MEPhI Program "Priority 2030".

I

Важнейшим источником, позволяющим получать информацию по истории литературного языка Древней Руси и других славянских народов, являются так называемые памятники традиционного содержания. Речь, в первую очередь, идет о богослужебных книгах, которые распространялись в огромном количестве списков. Теоретически в каждом храме должен был быть комплект таких книг. И хотя на практике это, конечно же, было не совсем так, в любом случае число книг, входящих в основной богослужебный круг (Служебник, Минея, Постная и Цветная Триодь, Паримейник, Служебное Евангелие, Следованная Псалтырь и др.), было очень большим. Поскольку эти книги переписывались в разных регионах, они содержат информацию об особенностях региональных традиций и книжных центров. В связи с новыми канонизациями, изменениями церковного устава и другими событиями эти книги дополнялись новыми текстами. А поскольку число сохранившихся рукописей измеряется многими десятками, а чаще сотнями, историки языка могут на их основе получить важную информацию.

Однако полноценно использовать материал рукописей традиционного содержания бывает достаточно сложно. Это связано с тем, что для них весьма затруднительно написать текстологическую историю, разбить на редакции и построить стемму[1], показывающую, как рукописи соотносятся друг с другом. Запутанность текстологической истории этих памятников связана с особенностями их переписывания. Переписчики богослужебных книг отдавали себе отчет в том, что в процессе копирования рукописей количество ошибок будет постоянно увеличиваться, поскольку

[1] В текстологии стеммой называется графическое изображение истории рукописей и взаимосвязей списков между собой.

при каждом следующем переписывании к ошибкам предыдущих писцов будут добавляться новые. Для того чтобы избежать подобной порчи текста, писцы, работая над рукописью, пользовались двумя оригиналами (текстологи называют их «антиграфами»). Переписывая один оригинал, писец попутно проверял себя по другому. Пока тексты совпадали, проблем не возникало. При обнаружении разночтения писец сравнивал между собой оба варианта и выбирал тот, который казался ему правильным [Алексеев 2001: 691–694]. Таким образом, переписывание книг с использованием двух антиграфов способствовало стабилизации текста. Случайные описки и пропуски не распространялись в процессе переписывания, а фиксировались и исключались. Такой способ бытования и копирования книг называют контролируемой традицией.

Исследователи, работающие с рукописями контролируемой традиции, сталкиваются с рядом трудноразрешимых задач. Дело в том, что методы классической текстологии создавались, в первую очередь, для исследования литературных текстов, при переписывании которых два антиграфа не используются. В этом случае для группировки рукописей бывает достаточно на основе анализа разночтений установить, что редакция А характеризуется одним набором признаков, а редакция В — другим. Понятно, что такая идеальная картина выстраивается далеко не всегда. Часто появляются разного рода смешанные редакции. Тем не менее изучение истории текста начинается с выявления сохранившихся рукописей и подведения разночтений, позволяющих выделить основные редакции данного текста, установить, кем, когда и для чего это редактирование проводилось, а результаты такого анализа часто представляют в виде стеммы, наглядно показывающей историю текста[2].

Когда же мы обращаемся к памятникам, принадлежащим к контролируемой традиции, методы классической текстологии перестают работать. Использование писцами двух антиграфов приводит к тому, что рукописи не имеют понятных наборов признаков, позволяющих выделить редакции. Текстовые и языковые особенности, возникшие в результате сознательного редактирования или исправления по греческому оригиналу, очень быстро распыляются по другим рукописям. Текст новой редакции оказывается одним из антиграфов, а писец, натолкнувшись на разночтение, в одном случае предпочтет старый вариант, а в другом — новый. В результате особенности разных редакций причудливо переплетаются. Поскольку тот набор рукописей, который до нас дошел, с точки зрения истории текстов является произвольным (сохранилось то, что не сгорело, не сгнило и не было съедено мышами), задача становится почти

[2] Классическое описание работы текстолога, занимающегося древнерусской письменностью, принадлежит Д. С. Лихачеву [Лихачев 1981: 175–244, 457–461].

неразрешимой. Из тысяч рукописей со случайно распределенными текстологическими особенностями произвольно взято небольшое количество экземпляров, на основании которых исследователи пытаются делать какие-то выводы.

Остановимся чуть подробнее на том, какая же часть от общего количества рукописей дошла до наших дней. В свое время Б. В. Сапунов попытался выяснить, сколько книг было необходимо для того, чтобы в древнерусских церквях совершалось богослужение. По его расчетам на Руси в XI–XIII вв. было построено около 10 тыс. церковных зданий. Можно составить минимальный список книг, без которых невозможно совершать богослужение. Значит, должно было быть изготовлено порядка десяти тысяч рукописных копий каждой из книг, входящих в этот список [Сапунов 1978: 64]. Посмотрим хотя бы для некоторых из этих книг, сколько экземпляров (включая малоинформативные отрывки) фиксирует Сводный каталог славяно-русских рукописных книг XI–XIII вв. [Сводный каталог 1984: 375–376].

Тип богослужебной книги	**Количество известных экземпляров**
Служебное Евангелие разных типов	116
Паремийник	12
Триодь постная	10
Триодь цветная	8
Триодь постная и цветная в одном томе	14

Само собой разумеется, эти цифры не претендуют на точность, и тому имеется много причин. Во-первых, сводный каталог фиксирует только рукописи, хранящиеся на территории бывшего СССР. Во-вторых, в числе рукописей каталога немало тех, которые были переписаны у южных славян. Поэтому, приводя эти цифры, мы имеем основания лишь утверждать, что до нас дошла ничтожная часть от переписанных в Древней Руси рукописей, принадлежащих к контролируемой текстологической традиции. Построить внятную текстологию, опираясь на разрозненные рукописи, в которых из-за использования нескольких антиграфов отсутствуют отчетливые текстологические приметы, едва ли возможно.

II

В конце 60-х годов XX века благодаря появлению компьютерных технологий стали создаваться методики исследования памятников с контролируемой традицией, построенные на иных основаниях. В 1969 г. появилась книга Э. Колвелла, посвященная анализу греческого текста Священного

Писания [Колвелл 1969]. А начиная с середины 80-х годов А. А. Алексеев и его школа разработали методику анализа славянских памятников контролируемой рукописной традиции. Вкратце она устроена так. Исследователь вручную выделяет по всем имеющимся в его распоряжении спискам те места, в которых встречаются разночтения. Такие вариативные фрагменты текста, называемые «узлами разночтений» (англ. Variation units), нумеруются. Варианты внутри каждого узла разночтения также нумеруются, и в результате текстологические особенности каждой рукописи оказываются описанными при помощи набора чисел. Затем компьютер попарно сравнивает все рукописи. Процитируем фрагмент описания этой методики: «ЭВМ рассматривает пары списков в порядке убывания процента общих чтений. Две рукописи, имеющие наибольшую степень сходства, образуют первый кластер. Затем рассматривается следующая пара рукописей: если одна из них уже вошла в первый кластер, то к нему приписывается и вторая; если обе рукописи "новые", то они объединяются во второй кластер. Таким образом, рукописи либо приписываются к уже существующим кластерам, либо образуют новые кластеры» [Алексеев, Кузнецова 1988: 115]. С помощью данной методики удается выделить группы рукописей, наиболее близких друг к другу в текстологическом отношении. Эта информация очень полезна при подготовке изданий, поскольку помогает выбрать как основную рукопись, так и рукописи, по которым будут подводиться разночтения. Однако описать историю исследуемого текста или соотнести редакции с той или иной выделенной группой рукописей, как правило, не удается. Это связано с тем, что статистическая близость рукописей не означает их генетической близости. Выделенные группы объединяются не идеей редактора, вносившего в текст изменения, а обезличенной статистической близостью. Характерно, что наиболее стабильными оказываются не древнейшие рукописи, стоящие у истоков бытования памятника, а более поздние, к моменту создания которых процесс стабилизации и унификации шел уже достаточно долго [Алексеев 2001: 696].

Применение методики Колвелла очень много дало исследователям истории текстов контролируемой традиции. В течение времени, прошедшего с момента появления программной статьи А. А. Алексеева [Алексеев 1985], была проведена серия важных исследований и публикаций, посвященных текстам славянских книг Священного Писания[3]. Результаты этих исследований славянской Библии были обобщены в монографии А. А. Алексеева [1999], а методика исследования текстов контролируемой

[3] Отметим серию работ А. А. Пичхадзе, завершившуюся публикацией паримейной версии книги Исход [Пичхадзе 1998], публикацию паримейной версии книги Иова [Афанасьева и Шварц 1980] и, наконец, фундаментальное издание под руководством А. А. Алексеева славянского текста Евангелия от Иоанна и Евангелия от Матфея с привлечением материала более чем тысячи рукописей.

традиции была, так сказать, канонизирована тем, что в третье издание классической «Текстологии» Д. С. Лихачева вошла подготовленная А. А. Алексеевым дополнительная глава, посвященная методам исследования текстов славянского Священного Писания [Алексеев 2001].

III

Описанная выше методика анализа рукописей контролируемой традиции чрезвычайно трудоемка, поскольку на начальном этапе исследователям приходится вручную выделять узлы разночтений в десятках, а то и в сотнях рукописей. Поэтому работ, авторы которых смогли завершить весь исследовательский цикл — от выделения узлов разночтений до критического издания текста по большому количеству списков, очень немного. Между тем за период, прошедший со времени появления книги Колвелла и программной статьи А. А. Алексеева, компьютерные технологии претерпели существенные изменения.

В 60-е годы, когда эти методы начали разрабатываться, компьютеры не могли работать со средневековыми текстами, поскольку существовавшие тогда кодировки не давали возможности работать со знаками, отсутствовавшими в современных алфавитах. С появлением Юникода эта проблема в принципе стала разрешимой. Появление технологий, позволяющих обрабатывать большие данные, и развитие корпусной лингвистики привели к появлению таких ресурсов, как исторические корпуса в составе Национального корпуса русского языка (ruscorpora.ru); проект «Манускриптъ» (manuscripts.ru), «Древнерусские берестяные грамоты» (gramoty.ru) и др., позволяющих исследователю осуществлять поиск по заданным параметрам по большому корпусу древних текстов. Однако в существующих корпусах инструментарий для проведения сравнительного анализа текстов рукописей по фрагментам либо ограничен, либо и вовсе отсутствует.

Благодаря современным технологическим достижениям появилась возможность поставить вопрос о развитии новых методов работы с текстами контролируемой традиции. Теперь нет необходимости вручную выделять узлы разночтений и нумеровать варианты, поскольку машиночитаемый текст средневековых рукописей доступен обработке при помощи программного инструментария. Разработка такого ресурса для работы с текстами древнерусских рукописей (на данный момент он называется «Лингвистическая интеллектуальная среда», или ЛИС) сейчас ведется в Лаборатории цифровой лингвистики МИФИ. Этот междисциплинарный проект осуществляется совместно с филологами, сотрудниками Института русского языка им. В. В. Виноградова РАН.

Первым этапом и основным условием реализации проекта должна стать программа автоматического распознавания рукописей, работа над которой ведется в настоящее время. Работающая программа распознавания — необходимое условие реализации проекта. Прецеденты создания программ автоматического распознавания рукописного текста нейросетевыми моделями существуют, см. программу Transcribus (transcribes.eu) и ее использование для автоматического распознавания церковнославянских рукописей [Rabus 2019]. Однако подобного отечественного ресурса до сих пор не существует, в то же время его создание актуально и востребовано, в особенности с учетом большого количества рукописей, уже оцифрованных архивами.

Работающая программа по распознаванию рукописного текста — ключевая черта, отличающая настоящий проект от других существующих корпусных проектов. Такая программа позволит создать объемный корпус и включить в него тексты рукописей, которые еще не были распознаны и в настоящий момент доступны для исследователей только в виде картинок на различных платформах. В данном случае предпочтение будет отдаваться максимально широкому охвату рукописей в ущерб идеально чистому тексту, поскольку машинное распознавание практически никогда не позволяет получить на выходе чистый текст, не требующий последующего ручного редактирования. Для решения данной проблемы предполагается со временем создать пользовательский модуль, позволяющий исследователю дополнительно натренировать программу на более чистое распознание конкретной выбранной рукописи. Благодаря такой возможности можно будет постепенно точечно улучшать качество распознанных текстов.

IV

Ниже мы попробуем обозначить те филологические и текстологические задачи, которые можно решить, используя современные технологии. Мы начнем с наиболее очевидной задачи — подготовки электронного издания средневековых богослужебных книг (в первую очередь, Служебных Миней и Триодей). В чем особенность этих книг и почему именно с них следует начать обкатку электронной системы исследования средневековых рукописей?

Дело в том, что эти книги членятся на небольшие по объему фрагменты (тропари, стихиры, библейские чтения, богослужебные указания, надписания и т. д.). Каждый из этих текстов может иметь свою историю: переходить из одной службы в другую, редактироваться, переосмысляться. Кроме того, позднейшие гимнографы часто использовали фрагменты

старых служб или законченные песнопения при составлении новых произведений. Приведем лишь один пример подобного заимствования. Третья стихира на «Господи, воззвах» службы Феодосию Печерскому (+1074), которая относится еще к домонгольскому периоду, представляет собой незначительную переделку стихиры, заимствованной из переведенной с греческого службы Феодосию Великому (ок. 424–529), основателю палестинского монашества.

Служба Феодосию Великому, 11 января, вечерня, 2-я стихира на «Господи, воззвах»[4]	Служба Феодосию Печерскому, 3 мая, великая вечерня, 3-я стихира на «Господи, воззвах»
Преподобне отче Богоносе Феодосие, обретши, якоже желаше, чистую твою душу, Духа благодать Всесвятаго в тя вселися, яко Пречистый Свет, Егоже действом светло украшен, Христа непрестанно славословиши, во двою существу Единаго Сына, крещаемаго рукою Предтечевою, и свидетельствована гласом Отчим. Того моли, Тому помолися, преподобне, даровати вселенней единомыслие, мир и велию милость [Минея 1983: 372]	Преподобне отче Богоносе Феодосие, обрел, якоже возлюбил еси, чистую твою душу, благодать бо Всесвятаго Духа в тебе вселися, яко пречист свет, Егоже поспешением светло озарен, Христа непрестанно славословил еси, во двою существу единаго Бога, пострадавшего плотию на Кресте и Божеством безстрастна пребывша. Того моли, Тому помолися, преподобне, даровати вселенней мир и велию милость [Минея 1987: 132]

Это далеко не единственное заимствование, встречающееся в службе преподобному Феодосию [Спасский 2008: 76–78]. Заимствований такого рода в богослужебных книгах очень много. Но для того чтобы проследить судьбу одного конкретного песнопения, исследователь должен обнаружить интересующий его текст во многих десятках рукописей и посмотреть, нет ли в нем заимствований из других текстов (для этого теоретически следует просмотреть все дошедшие до нашего времени источники). Понятно, что подобная задача является настолько трудоемкой, что решать ее методами традиционной текстологии никто никогда не будет. Но когда основной массив текстов будет оцифрован, появится техническая возможность сравнивать отдельные песнопения друг с другом.

[4] Для простоты мы цитируем эти тексты по современному обиходному изданию.

Задача поиска заимствований, обнаружения похожих или тождественных текстов уже давно решена в связи с разработкой инструментов для борьбы с плагиатом. Таким образом, после того, как будет решена задача относительно чистой оцифровки средневековых рукописей, начнется подготовка инструмента, позволяющего производить сравнение текстов с опорой на различные параметры.

Этот инструмент должен работать так. Оцифрованный текст разбивается на структурные элементы, причем каждый из них получает уникальный номер (или же имя). Далее машина сравнивает этот текст со всеми другими текстами, имеющимися в оцифрованном корпусе. В результате будут найдены все случаи вхождения данного текста в богослужебные книги.

Для исследователя процесс работы с ЛИС будет выглядеть следующим образом. Найдя интересующий его фрагмент, пользователь сможет посмотреть, в каких песнопениях этот фрагмент встречается. Далее он сможет обратиться к этим песнопениям и посмотреть, есть ли какие-то изменения, а если есть, то какие. Таким образом, в распоряжении исследователя окажется не только коллекция вхождений интересующего его слова или выражения, но и выборка слов или конструкций, появившихся на их месте в результате редактуры разного времени.

Фактически можно говорить о новом типе издания памятников с привлечением максимально широкого (в идеале — исчерпывающего) количества рукописей. На этом моменте следует остановиться подробнее. При подготовке привычных нам бумажных изданий памятников, сохранившихся во многих сотнях списков, текстологи должны решить практически нерешаемую задачу. С одной стороны, они стремятся привести максимальное количество разночтений. С другой — огромный справочный аппарат делает издание чрезвычайно громоздким и затрудняет работу с ним. В итоге на предварительном этапе приходится производить отбор рукописей, часть из них объявляя второстепенными. Точно так же приходится ограничивать и количество разночтений, которые будут приводиться в аппарате. В подавляющем большинстве изданий не фиксируются, например, орфографические разночтения, поскольку они не являются текстологически значимыми. Издания, аппарат которых позволяет полностью реконструировать все особенности рукописей, использованных при подготовке этого издания, представляют собой редчайшие исключения.

Электронное издание книг контролируемой традиции даст возможность приводить для каждого текста максимальное количество разночтений. При этом пользователь получит издание текста, максимально приближенное к его исследовательским задачам. Если его интересуют древнейшая версия текста, в качестве основного будет использоваться одна

из древнейших рукописей, если его интересует более поздняя эпоха, то будет выбрана и соответствующая рукопись. Таким образом, у исследователя появится возможность работать с электронным изданием, ориентированным на ту проблему, которой он занимается. Напомним, что та картинка, которую пользователь видит на экране, формируется автоматически на основе машинного анализа текстов, принадлежащих разным эпохам.

Следует особо отметить, что к числу возможностей создаваемого электронного инструмента относится и возможность подгрузки текстов на других языках. В первую очередь, это актуально для исследования переводов с греческого. Подгрузка греческого оригинала, а в идеале — электронного издания греческого текста — обещает появление полезнейшего ресурса. Но это уже дело будущего.

Отдельно следует сказать о возможности полной фиксации всей текстовой традиции, а не исключительно рукописной. Традиционно исследователи обращаются к древнейшим рукописям контролируемой традиции. Поздние рукописи исследованы намного хуже, чем древнейшие, а судьба тех же самых текстов в период книгопечатанья исследована в еще меньшей степени. Между тем единственным видом славянских текстов, которые на протяжении тысячелетия изменялись, но оставались тождественными сами себе, являются богослужебные книги. И электронная форма публикации книг контролируемой традиции дает возможность представить каждый из текстов во всех известных видах: от наиболее ранних редакций, читающихся в средневековых рукописях, до наиболее поздних, печатающихся в богослужебных книгах первой четверти XXI века.

V

В процессе распознавания рукописей будет создан огромный банк данных, содержащий варианты написания букв славянского алфавита, лигатур и надстрочных знаков. Создание подобного банка — рутинная операция, необходимая для того, чтобы обучить машину распознавать образы. Однако такой банк имеет большую ценность и сам по себе. Мы получаем возможность сформировать запрос типа «как выглядит буква А в датированных рукописях XII» или же в рукописях, переписанных, например, в Новгороде. На основании этого можно составить качественные полиграфические таблицы (в электронном или бумажном виде). К тому же, и это главное, имея образы букв, извлеченных из датированных рукописей, машина сможет ответить на вопрос, к начеркам какого времени ближе всего стоит рукопись, время создания которой неизвестно. Таким образом, появится возможность датировать рукописи по

палеографическим признакам на основании машинного анализа большого количества рукописей. Точно так же возможен автоматический анализ рукописей, место создания которых неизвестно, и сравнение их почерков с почерками тех рукописей, для которых надежно устанавливается место создания. Теоретически и та, и другая задачи вполне разрешимы, хотя понятно, что на этапе их непосредственной реализации нам придется столкнуться со многими трудностями. Подобного ресурса, включающего не только корпус древнерусских рукописей с разными почерками, но также и доступный пользователю автоматический анализатор, позволяющий сравнивать начерки отдельных букв внутри одной рукописи, а также почерки разных рукописей, на данный момент не существует, однако он несомненно был бы востребован исследователями.

VI

В предыдущих разделах мы говорили о тех возможностях ЛИС, которые прежде никогда не реализовывались. В завершение же статьи следует кратко назвать те очевидные функции, возможность использования которых несомненна и не представляет собой чего-то принципиально нового. Предполагается, что функционал ЛИС будет располагать стандартными возможностями, которые имеют существующие лингвистические корпусы (например, НКРЯ, Манускриптъ). В первую очередь, речь идет о различных видах поиска (по леммам, комбинациям грамматических признаков, сочетаниям букв, морфологическим элементам и т. д.), возможности сортировки результатов и их экспорта в Ecxel, Word и т. д. Существенной особенностью нашей системы является то, что для каждого примера можно будет легко перейти на страницу рукописи, откуда этот пример заимствован (такая функция не реализована в НКРЯ, в корпусе «Манускриптъ» она реализована частично). Этот момент нам представляется очень важным. Когда мы имеем дело со сложными в орфографическом отношении средневековыми рукописями, ссылки типа «цитируется по ЛИС» не являются удачными, поскольку распознанный текст будет постоянно улучшаться и вычитываться, а значит, несколько меняться. Поэтому, работая с системой, исследователь всегда имеет возможность обратиться к первоисточнику цитаты и сослаться на рукопись. К тому же необходимо отдавать себе отчет в том, что при масштабном машинном распознавании неизбежно появление значительного количества ошибок и идеально чистый текст получить не удастся. Поэтому у исследователя должна быть возможность проверки результатов. Наличие кнопки «сообщить об ошибке» позволит исправлять замеченные пользователями ошибки распознавания.

* * *

Развитие технологий дает возможность создания принципиально новых инструментов для гуманитариев, в частности для работы с так называемыми памятниками традиционного содержания. Технологии обработки больших данных позволяют осуществлять издания памятников по сколь угодно большому числу списков. При этом возникает новый тип издания памятника, которое не имеет устойчивого, навсегда заданного текста, а организует материал в соответствии с пользовательским запросом. Таким образом, технологии будущего дают новые возможности для работы с текстами далекого прошлого.

Источники

Минея 1983 — Минея. Январь. Ч. 1. М.: Издательство Московской Патриархии, 1983. 592 с.

Минея 1987 — Минея. Май. Ч. 1. М.: Издательство Московской Патриархии, 1983. 488 с.

Сводный каталог 1984 — Сводный каталог славяно-русских рукописных книг, хранящихся в СССР / Гл. ред. С. О. Шмидт. М.: Наука, 1984. 406 с.

Литература

Алексеев А. А. Проект текстологического исследования Кирилло-Мефодиевского перевода Евангелия // Советское славяноведение. 1985. № 1 С. 82–95.

Алексеев А. А. Текстология славянской Библии. СПб: Дмитрий Буланин, 1999. 255 с.

Алексеев А. А. Текстология переводных произведений (Священное Писание) // Лихачев Д. С. при участии Алексеева А. А. и Боброва А. Г. Текстология. На материале русской литературы XI–XVII в. СПб.: Алетейя, 2001. С. 689–717.

Алексеев А. А, Кузнецова Е. Л. ЭВМ и проблемы текстологии славянских текстов // Лингвистические задачи и проблемы обработки данных на ЭВМ / Под ред. Ю. Н. Караулова. М.: АН СССР, Институт русского языка, 1988. С. 111–120.

Афанасьева Е. В., Шварц Е. М. Древнейший славянский перевод книги Иова (по пергаменным рукописям) // Источниковедение литературы Древней Руси / Под ред. Д. С. Лихачева. Л.: Наука, 1980. С. 7–32.

Лихачев Д. С. Текстология. На материале русской литературы X–XI веков. Л.: Наука, 1983. 639 с.

Пичхадзе А. А. Книга «Исход» в славянском паримейнике. В сб.: Ученые записки Российского православного университета ап. Иоанна Богослова / Под ред. игумена Иоанна (Экономцева). М.: Российский православный университет, 1998. С. 5–60.

Сапунов Б. В. Книга в России в XI–XIII вв. Л.: Наука, 1978. 231 с.

Спасский Ф. Г. Русское литургическое творчество. М.: Издательский Совет Русской Православной Церкви, 2008. 544 с.

Colwell E. C. Studies in Methodology in Textual Criticism of the New Testament. Leiden: E. J. Brill, 1969. 175 p.

Rabus A. Recognizing handwritten text in Slavic Manuscripts: a neural-network approach using Transcribus // Scripta & e-Scripta. 2019. Vol. 19. P. 9–32.

References

Alexeev A. A. [The project of textual research of the Cyril and Methodius translation of the Gospel]. *Sovetskoe slavyanovedenie*, 1985, no. 1, pp. 82–95. (In Russ.)

Alexeev A. A. *Tekstologiya slavyanskoi Biblii* [Text History of the Slavonic Bible]. St. Petersburg, Dmitrii Bulanin Publ., 1999. 255 p.

Alexeev A. A. [Text History of Translated Works (Holy Scripture)]. Likhachev D. S. with the participation of Alekseev A. A. i Bobrov A. G. *Tekstologiia. Na materiale russkoi literatury XI–XVII v.* [Textology. Based on Russian literature 11th — 17th c.] St. Petersburg, Aleteiia Publ., 2001, pp. 689–717. (In Russ.)

Alekseev A. A., Kuznetsova E. L. [Computers and Problems of Text History of Slavic texts]. *Lingvisticheskie zadachi i problemy obrabotki dannykh na EVM* [Linguistic problems and problems of data processing on computers]. Ed. by Yu. N. Karaulov. Moscow, AS USSR Publ., 1988, pp. 111–120. (In Russ.)

Afanas'eva E. V., Shvarts E. M. [The oldest Slavic translation of the Book of Job (based on parchment manuscripts)]. *Istochnikovedenie literatury Drevnei Rusi* [Source study of literature of Ancient Russia] Ed. by D. S. Likhachev. Leningrad, Nauka Publ., 1980, pp. 7–32. (In Russ.)

Colwell E. C. *Studies in Methodology in Textual Criticism of the New Testament*. Leiden, E. J. Brill Publ., 1969. 175 p.

Likhachev D. S. *Tekstologiya. Na materiale russkoi literatury X–XI vekov* [Textology. Based on the material of Russian literature of the 10th — 11th centuries]. Leningrad, Nauka Publ., 1983. 639 p.

Pichkhadze A. A. [The book "Exodus" in the Slavic Lectionar]. *Uchenye zapiski Rossiiskogo pravoslavnogo universiteta ap. Ioanna Bogoslova* [Scientific notes of the Russian Orthodox University of the Apostle John the Theologian.]. Ed. by Igumen Ioann (Ekonomtsev). Moscow, Russian Orthodox University Publ., 1998, pp. 5–60. (In Russ.)

Rabus A. Recognizing handwritten text in Slavic Manuscripts: a neural-network approach using Transcribus. *Scripta & e-Scripta*, 2019, vol. 19, pp. 9–32. (In Eng.)

Sapunov B. V. *Kniga v Rossii v XI–XIII vv.* [The book in Russia in the 11th — 13th centuries]. Leningrad, Nauka Publ., 1978. 231 p.

Spasskii F. G. *Russkoe liturgicheskoe tvorchestvo* [Russian Liturgical creativity] Mockow, Publ. Council of the Russian Orthodox Church, 2008. 544 p.