О проекте виртуальной среды для исследования списков «Беседы трёх святителей»♣

© А.Г. Варфоломеев, М.Г. Бабалык, А.В. Пигин

Петрозаводский государственный университет
avarf@psu.karelia.ru

Аннотация

Рассматривается проект специализированной виртуальной среды для научных исследований, ориентированной на текстологический анализ списков литературных памятников вопросно-ответной формы, а именно – апокрифа «Беседа трёх святителей». Представлены история изучения списков, методы и алгоритмы построения стемм, существующие виртуальные среды для исследования текстов, а также архитектура проектируемого веб-приложения.

1 Введение

В допечатную эпоху, а иногда – и в более позднее время литературные произведения существовали в виде *списков* (копий, созданных переписчиками). Вариативность списков может быть различной: от незначительных механических ошибок, лексических замен, перестановки слов – до значительных редакционных отличий, свидетельствующих о творческом отношении писца к тексту. В связи с этим одной из самых трудных задач изучения рукописных литературных памятников оказывается текстологический анализ списков. Он состоит из попарного сравнения списков и редакций между собой, фиксации разночтений, моделирования вариантов изменения текстов, и, в итоге, построения так называемой *стеммы* – генеалогического дерева списков, представляющего историю развития литературного памятника [7].

Строго говоря, задача построения истинной стеммы является существенно недоопределенной в математическом смысле, так как исследователю не известны ни настоящие причины, ни возможные способы модификации списков. Поэтому в процессе текстологического анализа исследователь использует далеко не только информацию о степени формальной близости текстов, но и значительный объем экспертных знаний. Однако эти знания большей частью используются на этапе сравнения нескольких возможных стемм между собой. Само же создание стемм-«кандидатов» вполне может быть произведено алгоритмическим путем на основе формального сравнения списков между собой, если четко определить математическую модель текста списка, алгоритм нахождения расстояния между моделями и метод построения дерева на основе матрицы попарных расстояний. Поскольку модели, алгоритмы и методы могут быть разными, перед текстологами открывается широкое поле для компьютерных экспериментов.

С развитием веб-технологий перед представителями самых разных наук появилась заманчивая возможность перенесения компьютерных экспериментов со стационарных компьютеров в среду веб. Стали создаваться так называемые «виртуальные лаборатории», дополняемые сопутствующими ресурсами (словарями, тезаурусами, библиотеками). С широким распространением технологий совместного создания и редактирования контента пользователи виртуальных лабораторий получили возможность активно взаимодействовать между собой, образуя сетевые научные сообщества. Возникло и стало приобретать растущую популярность понятие «виртуальная среда для научных исследований» (virtual research environment, VRE) как один из главных компонентов e-Science и e-Humanities [12].

Компьютерные эксперименты по построению стемм оптимально подходят для реализации их в рамках виртуальной среды. Данная статья посвящена описанию проекта создания виртуальной среды для текстологического анализа многочисленных списков одного из известных апокрифических произведений – «Беседы трёх святителей» (далее – Беседа).

2 История изучения Беседы

Беседа – памятник греческого происхождения, возникший предположительно в V – VI вв. [9]. Апокриф построен в форме вопросов и ответов, изложенных от имени трех иерархов православной церкви IV века Василия Великого, Григория Богослова и Иоанна Златоуста. Считается, что первоначально памятник представлял собой вопросы и ответы религиозного содержания на темы ветхо- и новозаветной истории. Со временем состав Беседы пополнялся разнообразным материалом, чему способствовала форма диалога. В переводе на славян-

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

ский язык Беседа известна уже в XI веке [10]. Самые ранние русские списки датируются XV веком, до этого времени на Руси Беседа известна лишь в выписках. На русской почве апокриф завоевал такую популярность, что переписывался вплоть до XX века, поэтому списков Беседы на сегодняшний день сохранилось очень много.

Научное изучение и издание древнерусских памятников апокрифической литературы начинается с середины XIX века. История изучения Беседы также начинается в это же время. Одним из первых обратил внимание на этот памятник в 1859 году Ф.И. Буслаев. Большой вклад в изучение Беседы внесли В.Н. Мочульский, И.Я. Порфирьев, Н.Ф. Красносельцев, А.Н. Веселовский, Р. Нахтигал. Но и сейчас еще остается множество нерешенных проблем – практически не изучена история русских списков апокрифа, не определены все его возможные источники.

В 1994 году была защищена кандидатская диссертация, посвященная Беседе. Ее автор – Е.А. Бучилина – попыталась продолжить текстологическое изучение списков Беседы, начатое В.Н. Мочульским. Материалом для диссертации послужили 3 латинских списка, 20 опубликованных греческих списков (исследовательница делит их на три вида – ветхозаветный, новозаветный и смешанный – и выделяет 10 основных циклов вопросов и ответов); 105 славяно-русских опубликованных списков и 36 новых (неопубликованных) рукописных списков из российских хранилищ. Русские списки исследовательница делит на три вида – близкий к сербскому, смешанный и особый.

Из современных исследований памятника следует также назвать работы болгарского ученого Анисавы Милтеновой, которая занимается текстологическим изучением апокрифа на материале южнославянских списков. Исследовательница приходит к выводу о важности территориальных исследований для создания наиболее полной картины развития истории текста. Особое внимание она уделяет комментированию тематических блоков вопросов и ответов, в частности, вопросов об Адаме и Еве. В 2004 г. в Софии была издана монография А. Милтеновой [16], подводящая итог ее многолетних исследований. Но, несмотря на то, что А. Милтенова активно занималась также вопросами электронных публикаций литературных памятников, информационная система по текстам Беседы до сих пор не создана. То же можно сказать и в целом о литературных произведениях вопросно-ответного жанра, широко распространенных в средние века в Западной и Восточной Европе (жанр Erotapokriseis) и активно изучаемых в наше время традиционными методами [14].

3 Компьютерные методы и алгоритмы в текстологическом анализе списков

Любой рукописный памятник вариативен, и для его изучения необходимо описание его литературной истории (истории движения текста). Списки апокрифа вариативны не только по тематическому составу и комбинации вопросов и ответов, но и по количеству этих вопросов. Так, некоторые списки представляют собой своего рода выписки из апокрифа одного или нескольких вопросов, некоторые же списки насчитывают свыше 100 вопросов и ответов. Такое разнообразие затрудняет текстологическое изучение памятника. Установить генетические связи между редакциями и вариантами памятника оказывается практически невозможным традиционными методами. Необходимо привлекать компьютерные технологии.

Начиная с 1960-х годов, компьютерные технологии применяются для текстологического анализа. Ярким примером являются работы Л.И. Бородкина и Л.В. Милова [1, 2], демонстрирующие метод автоматизации построения возможных стемм древнеславянского памятника «Закон Судный людем». За несколько десятилетий предложено много алгоритмов [19, 20, 21], большинство из них заимствовано из кладистики – биологической научной дисциплины, занимающейся построением эволюционных (филогенетических) деревьев. В основе «кладистических» методов исследования текстов – определение степени вариативности одного текста относительно другого, что можно интерпретировать как расстояние между текстами. Однако способов определения расстояния между текстами существует очень много. Прежде всего, сами тексты можно рассматривать как последовательности или множества слов или символов, можно учитывать только относительно редко встречающиеся слова или определенные части речи. Вместо текстов можно использовать вектора признаков или заменять их какими-либо графовыми структурами. Для каждой подобной математической модели текста можно предложить, в свою очередь, много различных процедур определения расстояния (правда, далеко не все такие «расстояния» будут обладать свойствами метрики). Если же тексты имеют иерархическую структуру, как, например, произведения вопросно-ответного жанра, то алгоритм определения степени вариативности должен одновременно учитывать и внешнюю структуру документа, и текст внутри структурных блоков. Кроме того, алгоритм в таком случае будет нуждаться в предварительной подготовке текстов – в частности, в выделении структуры. Для облегчения подготовки текстов к компьютерному текстологическому анализу, а также для совместного проведения анализа группой исследователей могут быть использованы виртуальные среды, ориентированные на анализ текстов.

4 Виртуальные среды для исследования литературных текстов

Историки и филологи, как правило, изучают коллекции источников самостоятельно, используя разнообразные методы их обработки, ставящие целью выделение информации из текстов. Но при

этом возникает проблема практической невоспроизводимости и непроверяемости результатов исследований. Большинство используемых документов не опубликовано, их оригиналы находятся в архивах, и для проведения прямой, основанной на источниках проверки результатов требуется столько же времени, что и на само исследование. Поэтому доверие к результатам исследований, основанных на архивных источниках, и принятие их научным сообществом базируются на авторитете исследователя, ожидаемости полученных результатов и применении общепринятых методик.

Современная филологическая наука уже давно вступила на путь создания полнотекстовых баз данных литературных произведений, доступных в интернете. С 1987 г. развивается и широко применяется формат разметки электронных публикаций литературных текстов TEI [22]. Коллективом под руководством проф. В.А. Баранова (Ижевск) разработана информационно-поисковая система «Манускрипт» [8], позволяющая вводить, редактировать, хранить и обрабатывать древнерусские тексты, а также выполнять поисковые запросы в окне веб-браузера. Большими аналитическими возможностями обладает информационная система Gramoty.ru [6], предоставляющая изображения и тексты Новгородских берестяных грамот.

Публикация коллекций источников вместе с методиками и результатами исследований, проведенных на их основе, способна изменить традицию и приблизить методологию историко-филологического исследования к стандартам точных наук. Однако, несмотря на то, что источники публикуются в интернете уже давно, только в последние годы стали появляться проекты, направленные на повышение объективности исследования за счет публикации источниковой базы, организации совместной работы или предоставления методик, инструментов и результатов исследования.

Примером среды для совместной работы с текстами является MONK [15]. Пользователи этой виртуальной среды имеют возможность проводить исследования коллекций текстов, размеченных в формате TEI, с помощью набора программных средств. Сначала тексты с помощью токенизации и лемматизации превращаются в последовательности слов, затем подсчитываются статистические характеристики текстов, служащие основой для их сравнения и классификации. Результаты исследований могут быть записаны как составные части проектов.

Более широкой функциональностью обладает разрабатываемая в Германии система TextGrid [11], ориентированная на историко-филологические исследования. Она позволяет работать не только с текстами, но и с их изображениями, предоставляя среду для полного цикла изучения рукописей. Для масштабного проекта «Монастериум», начатого немецкими историками совместно с коллегами из Австрии, Венгрии и других стран и посвященного созданию электронного архива документов из архивов монастырей Центральной Европы [17], в университете Кельна (Германия) разрабатывается специализированный редактор EditMom для совместной распределенной работы по оцифровке и ручному распознаванию текстов средневековых грамот [13]. В Петрозаводском государственном университете уже несколько лет идет работа над созданием алгоритмического и программного обеспечения, поддерживающего работу сетевых сообществ исследователей текстов [3, 4, 5].

Однако следует отметить, что в настоящее время нет ни одной виртуальной среды для научных исследований, которая была бы посвящена текстологическому анализу литературных памятников, и, в частности, произведений с вопросно-ответной структурой. Поэтому можно констатировать отсутствие отечественных и зарубежных аналогов для предлагаемого нами проекта.

Необходимо, на наш взгляд, также обратить внимание на современные виртуальные среды для экспериментальных исследований в других областях знаний, в частности, в биоинформатике. Так, в среде myExperiment [18] информация о компьютерном эксперименте сохраняется в виде «пакета», включающего в себя исходные данные, результаты эксперимента, а также workflow – формализованное описание методики эксперимента. Эти описания методик понятны компьютеру, и их можно использовать для автоматизации повторного проведения эксперимента.

5 Проект виртуальной среды для исследования списков Беседы

В нашем распоряжении находится более 80 русских списков Беседы XV – XX вв. из рукописей, хранящихся в Петрозаводске (НАРК, КГКМ, музей «Кижи»), Санкт-Петербурге (БАН, ИРЛИ), Москве (РГБ) и других городах. Основными задачами сейчас являются систематизация и дальнейшая обработка материала с целью выявления типологически, а если возможно, и генетически близких списков; тематическая систематизация всех имеющихся вопросов и ответов с целью определения новых источников апокрифа.

Для решения этих задач мы предлагаем создать специализированную информационную систему, работающую в среде веб и обладающую функциональностью виртуальной среды для научных исследований. В рамках такой системы электронные тексты всех рукописей коллекции должны составить базу данных. Ввод новых текстов в систему должен предусматривать их разбиение на структурные элементы «вопрос – ответ», которым (сразу или впоследствии) присваиваются определенные типы. Процесс соотнесения элементов типам может производиться вручную или в полуавтоматическом режиме (когда информационная система сама предлагает свои варианты наиболее подходящих, по ее мнению, типов на основе сравнения слов, входящих в тексты вопросов и ответов, с заранее определенными наборами ключевых слов, характеризующих типы). Компьютер также может выдать рекоменда-

цию ввести новый тип для данной пары «вопрос – ответ». Эксперт, пользуясь подобными рекомендациями, либо соглашается с ними, либо вручную выбирает свой тип. Но вполне возможным выглядит также автоматическое присваивание типа на основе некоторых правил, например, присутствия ключевых слов в текстовом содержимом элементов.

Для выделения структуры в текстах мы предлагаем использовать технологию XML. Вполне обоснованным выглядит в такой ситуации использование общепринятых в мировой практике схем XML-разметки на основе формата TEI. В рамках этого подхода электронные тексты всех рукописей коллекции должны быть размечены с помощью схемы разметки, состоящей в минимальном варианте из трех тэгов: элементы с текстовым содержимым ** ** (вопрос), ** ** (ответ) и их содержащий элемент **<p>** с атрибутами **n** (номер) и **type** (тип, определяемый исследователем) для пары «вопрос – ответ».

Пример фрагмента размеченного списка[i]:

```
<TEI>
<teiHeader type="text">
….
</teiHeader>
<text>
   <front>
      <head>
      Беседа триех святителей Василия Великаго, Григория Богослова, Иоанна Златоустаго, выписано ис патерика римскаго
      </head>
   </front>
   <body>
      <p n="1" type="">
         <div type="q">
         1. Григорий рече: Кто первы наречеся на земли?
         </div>
         <div type="a">
         Василей рече: Сатаниил наречеся первый, и причтен бы(с)ть Господем ко ангелом в десятый чин, за гордость же его наречен бысть Сатана и диавол, и свержен ангелом с небеси на землю прежде создания Адамля за чатыре дни.
         </div>
      </p>
      <p n="2" type="">
         <div type="q">
         2. Иоан рече: Что высота небесная и широта земная и глубина морская?
         </div>
         <div type="a">
         Василий рече: Отец, Сын и Святый Дух.
         </div>
      </p>
      …..
   </body>
</text>
</TEI>
```

Размеченные тексты могут сравниваться между собой различными способами в зависимости от выбора математической модели текста и алгоритма определения расстояния между моделями – с помощью сравнения последовательностей типов вопросов и ответов, сравнения текстов, находящихся внутри элементов **<p>** одинаковых типов. Такие сравнения можно реализовать с помощью скриптов (например, на языке PHP). В итоге появляется возможность разработки веб-приложения, ориентированного на текстологический анализ списков Беседы, но допускающего применение к любым другим спискам иерархической (в частности, вопросно-ответной) структуры.

В отношении функциональности и интерфейса мы предлагаем взять за основу виртуальные среды MONK и myExperiment. Так же, как в указанных средах, пользователь должен после регистрации иметь возможность создавать свои проекты (аналоги «пакетов» среды myExperiment), состоящие из коллекций текстов с метаинформацией, размеченных в соответствии со стандартом TEI P5, выбранных программных средств для вычисления расстояний между текстами и построения стеммы, описаний методики эксперимента в виде workflow в формате XML, а также полученных промежуточных и финальных результатов (матриц расстояний и стемм, тоже в формализованном виде). В качестве исходных данных проекта можно использовать уже готовые матрицы расстояний, взятые из выполненных ранее проектов.

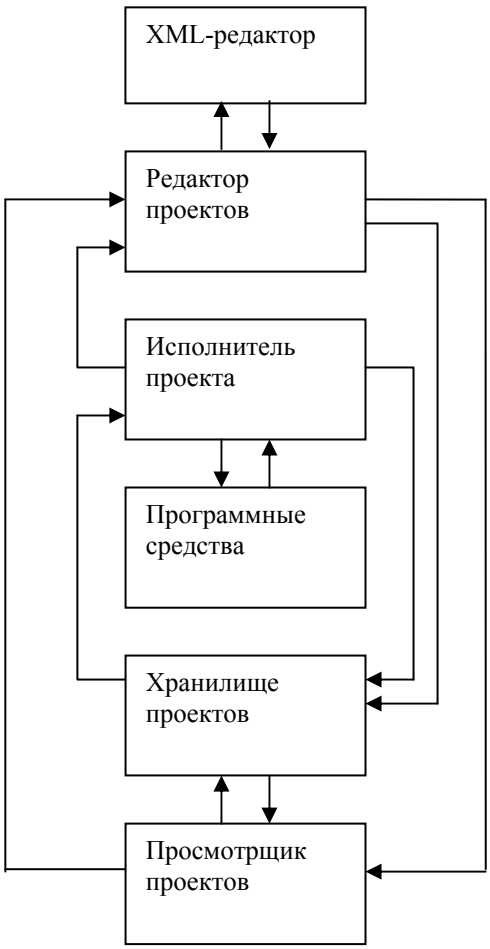
Трудным вопросом является степень открытости проектов для других участников сообщества. С одной стороны, исследователи могут быть не заинтересованы в доступе других к незавершенным проектам. С другой стороны, чем больше проектов будет открыто, тем эффективнее будет происходить обмен знаниями в сообществе. Поэтому мы считаем, что метаинформация о составе коллекции текстов и описание методики должны быть открыты у любого проекта. По завершении проекта исследователи должны получить доступ к его результатам. И, наконец, исходные данные (если только они не опубликованы библиотекой или архивом для всеобщего использования) должны быть доступны другим только при согласии их владельца.

Структурная схема работы виртуальной среды представлена на рисунке.

Заключение

В результате реализации проекта появится возможность проведения компьютерных экспериментов с различными алгоритмами классификации списков, и на их основе – автоматического построения наиболее вероятной стеммы списков. Разумеется, варианты стеммы, созданные компьютером, будут сильно зависеть от выбора правил присвоения вопросам и ответам определенных типов, сравнения списков и построения стемм, но в веб-ориентированной виртуальной среде можно обеспе-

чить сохранение правил (методик исследования) вместе с результатами для их дальнейшего использования, в том числе и для критики со стороны других исследователей. На наш взгляд, виртуальная среда с такой функциональностью позволит существенно повысить объективность текстологических исследований.



Литература

[1] Бородкин Л.И. Математические методы классификации древних текстов // Методы количественного анализа текстов нарративных источников. – М., 1983. – С. 8-30.

[2] Бородкин Л.И., Милов Л.В. О некоторых аспектах автоматизации текстологического исследования (Закон Судный людем) // Математические методы в историко-экономических и историко-культурных исследованиях. – М., 1977. – С. 230-279.

[3] Варфоломеев А.Г. Алгоритмическое и программное обеспечение работы сетевых сообществ исследователей текстов // Компьютерные науки и технологии. Ч. 1. Сб. трудов первой Межд. науч.-техн. конф. – Белгород: ГиК, 2009. – С. 30-33.

[4] Варфоломеев А.Г., Иванов А.С. Принципы электронных публикаций комплексов исторических документов со средствами палеографического, текстологического и дипломатического анализа // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. Материалы межд. науч. конф., Казань, 26 – 30 августа 2008 г. – Казань, 2008. – С. 60-63.

[5] Варфоломеев А.Г., Кравцов И.В. Аналитические Web-публикации исторических документов // Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет РФФИ: Труды Всерос. науч. конф., г. Новороссийск, 24 – 29 сентября 2007 г. – М.: Изд-во МГУ, 2007. – С. 389-390.

[6] Древнерусские берестяные грамоты. – http://gramoty.ru/.

[7] Лихачев Д.С. Текстология (на примере русской литературы X – XVII вв.). – М.-Л., 1962.

[8] Манускрипт. Древние славянские памятники. – http://manuscripts.ru/.

[9] Мочульский В.Н. Следы народной Библии в славянской и древнерусской письменности. – Одесса, 1893.

[10] Рождественская М.В. Беседа трех святителей (комментарии к тексту) // Памятники литературы Древней Руси. XII в. – М., 1980. – С. 649.

[11] Aschenbrenner A. et al. TextGrid – a modular platform for collaborative textual editing// Proc. of the Int. Workshop on Digital Library Goes e-Science (DLSci06). – Alicante, 2006. – P. 27-36.

[12] Blanke T. et al. Restful services for the e-Humanities – web services that work for the e-Humanities ecosystem// DEST '09. 3rd IEEE Int. Conf. on Digital Ecosystems and Technologies. – Istanbul, 2009. – P. 637-642.

[13] EditMom Editor. – http://www.editmom.unikoeln.de/.

[14] Erotapokriseis: Early Christian Question-and-Answer Literature in Context// Proc. of the Utrecht Colloquium, 13 – 14 October 2003. A. Volgers and C. Zamagni (eds.). – Leuven, 2004.

[15] Metadata Offer New Knowledge (MONK). – http://www.monkproject.org.

[16] Miltenova A. Erotapokriseis. Sachinenijata ot kratki v'prosi i otgovori v starobalgarskata literatura. – Sofia, 2004.

[17] Monasterium Project. – http://monasterium.net/.

[18] myExperiment. – http://www.myexperiment.org.

[19] O'Hara R.J., Robinson P.M.W. Computer-assisted methods of stemmatic analysis// Occasional Papers of the Canterbury Tales Project, Oxford, 1993. – V. 1. – P. 53-74. – http://rjohara.net/cv/1993-ctp/.

[20] Poole E. The computer in determining stemmatic relationships//Computers and the Humanities. 1974. – V. 8, No 4. – P. 207-216.

[21] Roos T., Heikkilä T.E. Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets//Literary and Linguistic Computing. – 2009. – V. 24, No 4. – P. 417-433.

[22] Text Encoding Initiative. – http://www.tei-c.org/.

The project of virtual research environment for textual criticism of "Conversation of the Three Holy Hierarchs"

A. Varfolomeyev, M. Babalyk, A. Pigin

In this paper an approach to organization of the virtual research environment for textual criticism of collections of literary manuscripts is examined. The approach allows saving not only the results but also the research methods in structured manner. Therefore the information about the research will be presented in a formalized form which allows the system to compare different researches and obtain new knowledge in the area of study.

The considered approach is used in practice in development of virtual research environment intended for textual criticism of question and answer literature, particularly "Conversation of the Three Holy Hierarchs".

♣ Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (проект 10-04-12118)

[i] Российская государственная библиотека. Фонд 214, № 241. Сборник поучений и сказаний, вторая половина XVIII века, скоропись