Филиппович Анна Юрьевна

Разработка словарных компонентов интегрированной информационной технологии переиздания печатных источников XVIII – нач. XIX вв.

Разработка проекта осуществляется при поддержке гранта Президента РФ № МК-3732.2010.9 в рамках конкурса молодых российских ученых-кандидатов наук 2010 года, направления "Информационно-телекоммуникационные системы и технологии".

На базе анализа печатных источников XVIII – нач. XIX вв., их палеографических и лексических характеристик и исследования современных технологий оцифровки книг (методик ввода и обработки текстов, OCR-систем, ИПС) разрабатываются словарные компоненты интегрированной информационной технологии переиздания на основе БД САР и осуществляется подготовка учебно-научных материалов.

Цели научного исследования

1. Создание словарных компонентов информационной технологии переиздания печатных источников XVIII – нач. XIX вв. для обеспечения эффективного доступа к ним научного сообщества на базе Словаря Академии Российской 1789-1794 гг.
2. Формирование в электронной форме лексического ядра языка коллекции документов рассматриваемого периода, лежащей в основе работы ИПС и OCR-систем.

Формулировка решаемой проблемы

Актуальной проблемой нашего общества является сохранение культурно-исторического наследия. Решение этой проблемы – перевод источников и архивных документов в цифровую форму. В своем выступлении на церемонии открытии Государственного исторического архива в Санкт-Петербурге 23 декабря 2005 г. Президент РФ В.В.Путин так охарактеризовал ее решение: «…это переезд и перенос имеющейся информации в документах на электронные носители, перенос в «цифру». Уверен, что если мы это сделаем – а мы сделаем это обязательно, – то это будет настоящая революция в изучении нашей отечественной истории, потому что позволит исследователям работать с любым документом» [http://www.kremlin.ru].

В результате активного использования современных информационных технологий и компьютерных методов обработки информации для сохранения и научного исследования древнерусского исторического наследия сложилась следующая проблемная ситуация. Основное внимание уделяется задаче сохранения исторических памятников. Решение проблемы хранения источников в электронной форме и их доставки потенциальным исследователям является, несомненно, первой

важнейшей задачей современности. Фактически осуществляется перенос исторических документов на новые носители, обеспечивающие более надежное и эффективное хранение только с точки зрения организации новых форм доступа исследователей к этим данным. Основным методом переноса на новые носители является оцифровка данных, подразумевающая факсимильное копирование источников. Копии, полученные таким образом, сопровождаются только библиографическими и археографическими описаниями. Основным недостатком такого типа описаний является неполное и ограниченное раскрытие содержания документа. В итоге, оцифровка источников фототипическим (факсимильным) способом принципиально не изменяет способ доступа к информации. По-прежнему ученый исследователь должен просматривать значительное количество источников для поиска нужной информации, последовательно "листая" их. В связи с этим, важнейшей второй задачей является решение проблемы поиска информации в созданных электронных хранилищах документов по их содержанию.

В системах хранения современных документов широко используются методы индексирования текста, позволяющие каждому документу поставить в соответствие его поисковый образ. Хранилища современных документов – это сами документы и их неотъемлемая часть автоматически (автоматизированно) полученные их поисковые образы. Это позволяет пользователям хранилищ документов указывать в своих поисковых предписаниях не только названия документов, даты их создания, авторов-создателей и т.п., но и конкретные факты, раскрывающие их содержание, что позволяет им на практике более полно удовлетворять свои поисковые потребности. Такие информационно-поисковые возможности для хранилищ исторических документов в настоящее время отсутствуют. Реализация их представляет собой актуальную научно-практическую проблему.

Первый вариант решения этой проблемы – осуществлять индексирование по мере изучения (прочтения и понимания) данных источников специалистами историками и филологами. Фактически в этом случае документ становится "общедоступным" только после того, как он освоен исследователем его "первооткрывателем". Иной вариант может быть основан на выделении группы источников с определенными свойствами (в числе которых способ печати, используемые средства оформления, шрифтовые гарнитуры и т.п.) и обработки их автоматизированными системами распознавания и индексирования текстов.

Основным компонентом таких систем является лингвистическая БД, ядро которой – лексическая система языка рассматриваемого исторического периода. Главная идея, лежащая в основе предлагаемого проекта – это формирование в электронной форме лексического ядра языка коллекции документов. Данную идею предлагается реализовать на одном из значительных культурно-исторических пластов, материале письменных источников XVIII – нач. XIX вв., которых по предварительной оценке только в фондах РГБ более тысячи.

В качестве основы ядра можно использовать Словарь Академии Российской 1789-1794 гг. (САР), содержащий более 200000 лексических

единиц. В 2001-2005 гг. Словарь Академии Российской был переиздан с использованием современных информационных технологий (руководитель проекта является разработчиком шрифтовой гарнитуры, используемой для набора текста и автором-дизайнером художественного оформления переиздания).

В течение 2006-2008 гг. в рамках проекта РГНФ "Интегрированная инструментальная информационно-программная среда для автоматизации исследований САР" были созданы: электронное издание САР, содержащее лингвистическую базу данных объемом более 44 тысяч структурных единиц и информационный ресурс (http://philippovich.ru/Projects/ESAR/ESAR.htm). Ресурс доступен для исследователей с 2007 года и содержит: PDF-издание, гиперграфическую систему факсимильных копий страниц оригинального 6-ти томного издания Словаря Академии Российской (объемом около 4000 страниц), электронный именной указатель переиздания. Посещаемость ресурса за последний год составляет более 120 тысяч пользователей и более 500 тысяч просмотренных страниц.

Задачи научного исследования

1. Анализ печатных источников XVIII – нач. XIX вв. и выявление их палеографических и лексических характеристик.
2. Исследование современных технологий оцифровки книг, методик ввода и обработки текстов и изображений, систем оптического распознавания, систем информационного поиска и автоматического индексирования документов.
3. Разработка словарных компонентов интегрированной технологии переиздания источников XVIII – нач. XIX вв. на основе БД Словаря Академии Российской 1789-1794 гг.
4. Разработка интегрированной технологии переиздания источников XVIII – нач. XIX вв. и исследование ее эффективности.
5. Подготовка учебно-научных материалов для исследования эффективности представленных технологий и подготовки ее отдельных компонентов.

Содержание исследования

1. Анализ печатных источников XVIII – нач. XIX вв. и выявление их палеографических и лексических характеристик включает решение следующих вопросов:
   - 1.1. Классификация печатных источников XVIII – нач. XIX вв. на основе их палеографических характеристик, к которым относятся особенности набора текста, формат изданий, качество бумаги, чернил и т.п. следы времени, используемые шрифтовые гарнитуры. Наиболее значительное влияние на эффективность технологии ввода текста оказывают последние. Использование ограниченного набора шрифтов в

рассматриваемый период времени позволяет выбрать типовые издания для оценки эффективности разрабатываемой информационной технологий переиздания.

- 1.2. Ввод фрагментов основных видов источников и исследование их лексических характеристик (сканирование страниц фрагментов выбранных источников).
- 1.3. Распознавание фрагментов страниц и предварительная оценка эффективности работы OCR-систем.
- 1.4. Формирование шрифтовых эталонов фрагментов, используя технологию обучения.
- 1.5. Квантитативные исследования текста, формирование словников фрагментов, построение функций распределения частот.

2. Исследование современных технологий оцифровки книг, методик ввода и обработки текстов и изображений, систем оптического распознавания, систем информационного поиска и автоматического индексирования документов включает решение следующих вопросов:
   - 2.1. Исследование моделей и технологий распознавания текстов с помощью OCR-систем, выявление характеристик, влияющих на эффективность технологий ввода текста.
   - 2.2. Исследование систем автоматического индексирования документов и информационно-поисковых систем, выявление характеристик, влияющих на эффективность поиска.
3. Разработка словарных компонент интегрированной технологии переиздания источников XVIII – нач. XIX вв. на основе БД Словаря Академии Российской 1789-1794 гг. включает решение следующих вопросов:
   - 3.1. Проверка и унификация данных для выявления ошибок и их устранения в БД, а также упрощения структур статей.
   - 3.2. Создание инструментария для создания словарных компонент – программ и скриптов.
   - 3.3. Расширение элементов БД Словаря Академии Российской 1789-1794 гг. за счет создания дополнительных таблиц и выборок.
   - 3.4. Формирование расширенного словника словоформ САР – основы лексического ядра языка XVIII – нач. XIX вв. за счет включения в словник заголовочных слов вариантов употребления слова.
   - 3.5. Создание базы цитатного материала для проектирования компонентов контекстной проверки текста на базе иллюстративного материала САР.

4. Разработка интегрированной технологии переиздания источников XVIII – нач. XIX вв. и исследование ее эффективности включает следующие вопросы:
   - 4.1. Исследование эффективности технологий ввода текста с помощью современной OCR-системы на базе созданных словарных компонентов и эталонов шрифтов.
   - 4.2. Сопоставительные исследования лексики введенных фрагментов источников XVIII – нач. XIX вв. и лексического ядра САР.
   - 4.3. Анализ систематических ошибок распознавания и формирование процедур их устранения.
   - 4.4. Индексация фрагментов основных видов источников для проектирования компонентов информационного поиска по источникам.
5. Подготовка учебно-научных материалов для исследования эффективности представленной технологии и подготовки ее отдельных компонентов.
   - 5.1. Разработка учебных программ и подготовка лекций по дисциплинам: «Информационные технологии исторической лексикографии», «Дизайн и оформление печатных изданий XVIII – нач. XIX вв.» (образовательная программа подготовки магистров «Компьютерная лингвистика и семиотика» для студентов Московского государственного университета печати – МГУП специальностей 230203 – «Информационные технологии в медиаиндустрии», 230204 – «Информационные технологии в дизайне», поток 20-30 чел.);
   - 5.2. Подготовка новых лекций по дисциплинам «Семиотика информационных технологий» и «Лингвистическое обеспечение АСОИУ» для студентов МГТУ им. Н.Э.Баумана (образовательная программа подготовки инженеров по специальности 230102 – «Автоматизированные системы обработки информации и управления» поток 90-110 чел.);
   - 5.3. Подготовка методических материалов курсовых проектов, домашних заданий и практических занятий для проведения исследований эффективности технологий ввода и информационного поиска текста на материалах источников XVIII – нач. XIX вв.
   - 5.4. Подготовка учебного пособия «Информационные технологии исторической лексикографии» и/или «Методы оцифровки и распознавания рукописных и первопечатных древнерусских источников» для изучения подходов к оцифровке исторических письменных источников.

- 5.5. Курсовое и дипломное проектирование, бакалаврские и магистерские квалификационные работы.
- 5.6. Написание не менее 8-ми статей и тезисов по теме проекта.
- 5.7. Формирование Учебно-научного студенческого коллектива из числа студентов МГУП и МГТУ им. Н.Э.Баумана.

Новизна научного исследования:

1. Впервые будут получены частотные и индексированные словники текстовых фрагментов источников XVIII – нач. XIX вв, выявлены параметры функции распределения частот слов и динамики появления новых слов.
2. Будет сформировано лексическое ядро языка XVIII – нач. XIX вв, для дальнейшего использования в качестве основной компоненты работы ИПС и OCR-систем.
3. Будет сделано теоретическое обоснование интегрированной информационной технологии переиздания печатных источников XVIII – нач. XIX вв.

Методы решения задач научного исследования:

Для решения поставленных задач используются методы компьютерной (квантитативной) лингвистики, математической статистики, исторической лексикологии и лексикографии; теории частот слов, множеств, вероятности, формальных языков и грамматик; объектно-ориентированные и реляционные подходы к проектированию ИС.

Ожидаемые результаты исследования:

В результате выполнения проекта будут получены новые научные данные на базе исследований источников XVIII – нач. XIX вв. Выявлены основные типы печатных источников XVIII – нач. XIX вв. описаны их палеографические и лексические характеристики, построена квантитативная модель текстов, включающая частотные и индексированные словники, параметры функции распределения частот слов и динамики появления новых слов.

Будет создано лексическое ядро коллекции документов XVIII – нач. XIX вв., позволяющее решить проблему поиска информации в электронных хранилищах данных. 3. Представлено теоретическое обоснование принципов создания систем информационного поиска по историческим документам, технологий ввода текста с помощью OCR-системы и созданных словарных компонент для переиздания источников XVIII – нач. XIX вв.

Основные направления дальнейшего использования предполагаемых результатов

Проект направлен на обеспечение информационным материалом и современным инструментарием ученых, занимающихся проблемами формирования и развития норм русского литературного языка. Разработка словарных компонентов интегрированной информационной технологии переиздания печатных источников XVIII – нач. XIX вв. повысит эффективность проведения лексикографических исследований на их материале, которые необходимы для активизации работы в области исторической лексикологии и лексикографии русского языка, исследования памятников письменности, составления словарей в области истории русского языка и литературы, древнерусского искусства.

Результаты проекта могут быть использованы для переиздания значительного массива конкретных источников XVIII –нач. XIX вв., разработки систем распознавания исторических текстов, решения практических задач электронного и полиграфического издания древних памятников.

На основе созданного лексического ядра и других словарных компонент могут быть разработаны информационно-поисковые системы, реализующие эффективный доступ к информации, хранящейся в источниках XVIII –нач. XIX вв.

Работы над проектом и основные его результаты будут иметь культурно-нравственное воспитательное значение для студентов технических специальностей высших учебных заведений и молодых исследователей.

На базе результатов проекта могут быть начаты научно-исследовательские работы по освоению других исторических пластов культурного письменного наследия.