Ю.Н. Филиппович, Г.А. Черкасова, М.И. Чернышева

ПРОЕКТ РОССИЙСКОГО ГУМАНИТАРНОГО НАУЧНОГО ФОНДА «ИНТЕГРИРОВАННАЯ ИНСТРУМЕНТАЛЬНАЯ ИНФОРМАЦИОННО-ПРОГРАММНАЯ СРЕДА ДЛЯ АВТОМАТИЗАЦИИ ИССЛЕДОВАНИЙ СЛОВАРЯ АКАДЕМИИ РОССИЙСКОЙ 1789–1794 гг.»

Научные проблемы, на решение которых направлен проект. Проект направлен на обеспечение решения проблемы формирования и развития норм русского литературного языка, которая впервые в истории русского языка была осознана и получила попытку многостороннего лексикографического решения в уникальном шести томном Словаре Академии Российской (САР), первое издание которого по корнесловному принципу было осуществлено в 1789–1794 гг.

Цель и задачи проекта — создание интегрированной инструментальной информационно-программной среды для автоматизации исследований Словаря Академии Российской 1789–1794 гг., которая позволит представлять процесс формирования и развития русского литературного языка и его тенденций.

Реализация проекта направлена на решение следующих основных научных задач:

1. Введение в научный оборот электронной версии САР;

2. Разработку модели построения несуществующего до настоящего момента словника САР;

3. Проведение лексико-семантического анализа помет, используемых внутри словаря, что позволит определить динамику развития взглядов на проблему формы русского литературного языка;

4. Создание частных словников для решения конкретных исследовательских задач, одной из которых является выявление смены подхода к отбору терминологической, специальной и заимствованной лексики при создании словарей;

5. Выявление источниковой базы САР.

Актуальность проекта обосновывается следующими положениями:

Во-первых, инструментальные средства, разрабатываемые в проекте, поддерживают актуальное научное направление и конкретные исследовательские задачи, впервые направленные на информационное

моделирование и лексико-семантический анализ уникального памятника русской лексикографии — первого толкового словаря русского языка — гнездового издания Словаря Академии Российской1789–1794 гг.

Во-вторых, разработка современных инструментальных средств исследования словарного материала XVIII века позволит привлечь внимание к данному направлению многих ученых, в том числе и молодых, а создание словарных баз данных и информационной системы обеспечит доступность «первого детища Академии Российской» для отечественных и зарубежных ученых, студентов и аспирантов.

В-третьих, компьютерная оснащенность области исследований истории языка существенно отстает от других областей лингвистики. Проект способствует сокращению существующего разрыва между количеством специальных информационных и программных средств современной и исторической компьютерной лингвистики.

В-четвертых, в настоящее время завершена допечатная (электронная) подготовка полиграфического переиздания всех шести томов Словаря Академии Российской 1789–1794 гг. Проект предлагается своевременно — вслед за завершающимися работами, а предлагаемые сроки его реализации позволяют с минимальными затратами создать базы данных и информационные системы.

Сфера использования проекта — это учебные и научные институты филологического профиля Министерства образования и науки РФ и РАН, Институт русского языка им. В.В. Виноградова.

Круг пользователей результатов проектных работ — отечественные и зарубежные исследователи, изучающие языковую культуру XVIII века, филологи, лексикографы, историки, издатели словарей, преподаватели, аспиранты, студенты.

Содержание проекта. В интегрированную инструментальную информационно-программную среду для автоматизации исследований Словаря Академии Российской 1789–1794 гг. входят шесть компонент:

1. Лингвистическая база данных.

2. Программный комплекс автоматизированного ввода текста словарных статей в базу данных.

3. Программный комплекс создания электронных словников (прямых и обратных, частотных, заголовочных слов, словоформ эксцерпций и дефиниций), словоуказателей (общих, предметных, именных), словарных подмножеств (языковых и тематических), поисковых тезаурусов.

4. Двухуровневая (запросная и гипертекстовая) информационно-поисковая система.

5. Биобиблиографическая информационная система «Создатели САР».

6. Гиперграфическая система факсимильной копии оригинального издания Словаря Академии Российской 1789-1794 гг.

Информационные системы предполагается реализовать в двух версиях: а) локального информационного ресурса на CD (DVD) дисках и б) Интернет-ресурса. Основным источником данных являются оригинальное и печатное переиздание Словаря Академии Российской 1789–1794 гг. т.1–6. М.: МГИ им.Е.Р.Дашковой, 2001–2005. (43257 словарных статей, около 200 авт. листов текста).

При решении задач проекта предлагается использовать методы и подходы исторической лексикологии и лексикографии, компьютерной лингвистики, теоретической и практической информатики. Будут применяться технологии локальных баз данных с использованием BDE (Borland Database Engine), SQL (Structured Query Language), технологии быстрой разработки приложений Delphi RAD (Rapid Application Development). В проекте предлагается использовать и апробированные информационные технологии, созданные ранее членами коллектива — уникальные алгоритмы сортировки, поиска, экспорта/импорта и хранения полиязычных текстов в СУБД. Разработанные информационные технологии и лексикографические методики подробно изложены в публикациях авторов проекта.

Общий план работ на весь срок выполнения проекта.

2006 год.

1. Разработка лингвистической базы данных.

2. Разработка программного комплекса автоматизированного ввода текста словарных статей в базу данных.

3. Подготовка и ввод в базы данных текста 1, 2 и 3 томов САР.

4. Разработка биобиблиографической информационной системы.

5. Разработка гиперграфической системы.

6. Подготовка изображений страниц 1, 2, 3 и 4 томов и ввод их в гиперграфическую систему.

7. Разработка программного комплекса создания электронных словников, словоуказателей, словарных подмножеств, поисковых тезаурусов.

2007 год.

1. Подготовка и ввод в базы данных текста 4, 5 и 6 томов САР.

2. Разработка двухуровневой (запросной и гипертекстовой) информационно-поисковой системы.

3. Подготовка изображений страниц 5 и 6 томов и ввод их в гиперграфическую систему.

4. Ввод данных в биобиблиографическую информационную систему.

2008 год.

1. Создание производных баз данных САР (вспомогательных и индексных баз словников, словоуказателей, словарных подмножеств и тезаурусов).

2. Разработка HELP-компонент Информационных систем, сопроводительной документации (научного, технического и методического сопровождения интегрированной среды), полиграфический дизайн (Picture-disk, упаковка и т.п.), Web—дизайн.

3. Комплексная отладка и интеграция компонент CD (DVD) диска САР.

4. Формирование и отладка Интернет ресурса.

Ожидаемые в конце 2006 г. результаты

В конце 2006 года предполагается получить следующие результаты:

1. Программный комплекс автоматизированного ввода текста словарных статей в базу данных.

2. Базы данных САР в объеме трех томов.

3. Рабочую версию программного комплекса создания электронных словников, словоуказателей, словарных подмножеств, поисковых тезаурусов по базе данных.

4. Сводный словник САР, содержащий словоформы из Показаний всех шести томов Словаря и их поисковые указатели (том, столбец), а также пометы (наличие/отсутствие в тексте словарной статьи, изменение, дополнение) и комментарии.

5. Гиперграфическая система факсимильной копии страниц оригинального издания САР в объеме четырех томов.

Современное состояние в данной области науки

Русская историческая лексикология и лексикография переживает новый этап своего развития: идет процесс перехода от традиционных

методов работы к современным, учитывающим, как достижения в области лексикографии современного русского языка (теоретическое обоснование и практическое создание разных типов словарей), так и возможностей компьютерной техники. Первым итогом нового этапа стал вышедший в издательстве «Наука» в 2001 г. Справочный том к Словарю Русского языка XI–XVII вв., содержащий указатель его источников и около 80 тыс. лексических единиц в обратном словнике, который охватывает 25 томов этого Словаря. Следующим шагом, позволяющим расширить исследовательскую базу в русской исторической лексикологии и лексикографии, может стать выпуск электронной версии САР.

Ввиду того, что САР в его первом издании представляет собой словопроизводный словарь двух языков «словенского» и «российского» со сложной многоступенчатой словарной структурой, то, как показывают наблюдения, в заголовочной строке и в словарных статьях находится больше словарных единиц, чем отражено в частных словниках, прилагаемых в виде Показаний к каждому тому. Такое положение дел приводит к тому, что: а) лексический массив САР недостаточно используется в лексикографической практике; б) бытует неточная оценка его количественного состава (43257 слов в 1-ом издании — по сведениям А.Красовского, 51388 слов во 2-ом издании — по подсчетам Е.Э.Биржаковой, где, как считается, полностью развернуты словообразовательные гнезда; в) делаются неверные выводы о соотношении лексического состава первого и второго изданий САР.

В настоящее время отсутствует в науке современный указатель источников САР.

Проведение лексикографических исследований на материале САР необходимы для активизации работы в области исторической лексикологии и лексикографии русского языка, исследования памятников письменности, в том числе переводных, исследования их языковых и литературных особенностей, решения практических задач электронного и полиграфического издания древних памятников, составления исторических словарей в области истории русского языка и литературы, древнерусского искусства, культуры, а также своевременного оказания практической помощи лексикографам, издающим Словарь русского языка XI–XVII вв., другие исторические словари.

Проект создания электронного издания такого значительного информационного ресурса — лексикографического памятника XVIII века Словаря Академии Российской 1789–1794 гг. является уникальным, а

его реализация будет достижением мирового уровня современной русистики.

Авторы проекта предполагают сначала разработать лингвистическую базу данных САР, затем программный комплекс ее автоматизированного наполнения, который позволит внести данные из файлов оригнал-макета печатного издания в поля базы данных. Используя ранее созданные программные средства, сформировать словники словоформ, частотные словники и конкордансы из файлов печатного издания. Параллельной работой является объединение Показаний томов САР и их филологический анализ. После создания базы данных САР предполагается сформировать производную базу источников САР и выполнить работы по ее расширению и уточнению современными библиографическими описаниями и филологическим комментарием. Независимо от выполняемых работ предполагается обрабатывать изображения страниц оригинала САР и формировать графическую базу данных факсимильной копии оригинального издания САР и программ работы пользователей с ней. Создание биобиблиографической справочной системы САР авторы предполагают выполнить на втором этапе работ, на завершающем этапе предполагается выполнить интеграцию всех баз данных, поисковых и запросных систем. После завершения всех намеченных работ предполагается формирование CD (DVD) дисков и Интернет ресурса.

Прямые аналоги проекта отсутствуют. Работы по созданию интегрированной инструментальной информационно-программной среды для автоматизации исследований Словаря Академии Российской 1789–1794 гг. ранее нигде не проводились. Базы данных Словаря Академии Российской 1789–1794 гг. в настоящее время не существует, а его электронные издания в форме CD (DVD) дисков и Интернет-ресурса будут введены в научный оборот впервые.

Близкими функциональными аналогами разрабатываемой в проекте интегрированной среды являются: «Система автоматизированного анализа естественно-языкового описания предметной области Интерлекс» (наиболее полное описание в книге Ю.Н.Филиппович, А.В.Прохоров. Семантика информационных технологий: опыты словарно-тезаурусного описания. С предисловием А.И.Новикова. М.: МГУП, 2002. — С.117–237.); «Информационно-поисковая система «Указатель источников» (наиболее полное описание в книге Ю.Н.Филиппович, А.Ю.Филиппович. Электронный указатель источников Рукописной древнерусской картотеки и Словаря русского языка XI–XVII вв. М.:

МГУП, 2002. — С.31–56.); экспериментальные разработки программ автоматизации создания и наполнения баз данных Словаря русского языка XI–XVII вв. (описаны в сборниках: Русская историческая лексикография на современном этапе. К 25-летию издания Словаря русского языка XI–XVII вв. / Отв. ред. Чернышева М.И. М.: ИРЯ РАН, 2000; Интеллектуальные технологии и системы. Вып. 3 / Сост. и ред. Ю.Н.Филиппович. М.: МГУП, 2001).

Имеющийся у коллектива научный задел по проекту

Научный коллектив имеет значительный научный и практический задел по данному проекту. В течение пятнадцати последних лет члены коллектива проводят исследования и разработки в области компьютерной лингвистики (в частности — исторической лексикологии и лексикографии). Работы проводятся по планам ИРЯ РАН, ИЯ РАН, а также по грантам РГНФ и РФФИ. Общее количество публикаций членов коллектива за этот период более сорока.

Члены коллектива являются основными участниками работ по переизданию Словаря Академии Российской (Филиппович Ю.Н. — руководство работами по переизданию, Чернышева М.И. — редактирование и корректура, Черкасова Г.А. — компьютерная верстка и дизайн, Филиппович А.Ю. — шрифтовое и компьютерное обеспечение, Немкова И.А. — корректура, Державина Е.И. — автор материалов о создателях Словаря). В настоящее время вышли из печати пять томов переиздания, заключительный шестой том выйдет до конца первого квартала 2006 года.

Коллектив располагает компьютерными версиями текста всех томов Словаря, разработанными оригинальными компьютерными шрифтами и компонентами графического материала. Для проведения работ по проекту может быть привлечен персонал, обученный для работы в среде СУБД Paradox 5.0, для автоматизированного заполнения словарных баз данных. Предполагается использовать апробированные ранее в других работах членов коллектива уникальные алгоритмы сортировки, поиска, экспорта/импорта и хранения полиязычных текстов в СУБД, а также технологические навыки и приемы обработки словарных данных.

Неполный список опубликованных работ коллектива

Ю.Н.Филиппович. О переиздании Словаря Академии Российской 1789–1794. Словарь Академии Российской 1789–1794. Том 1. М.:

МГИ им.Е.Р.Дашковой, 2001. С.7–10.

Ю.Н.Филиппович. Информационная технология электронного издания рукописных и первопечатных памятников древнерусской письменности / Издательское дело и редактирование: теория, методика, практика. Межведомственный сборник научных трудов. Вып.6. М.: МГУП, 2002. С.45–88.

Ю.Н.Филиппович, А.Ю.Филиппович. Электронный указатель источников Рукописной древнерусской картотеки и Словаря русского языка XI–XVII вв. М.: МГУП, 2002. 423 с.

Ю.Н.Филиппович, М.И.Чернышева. Историческая лексикография — terra incognita в компьютерном мире. Компьютерра, 1999, № 45. С.

Ю.Н.Филиппович, М.И.Чернышева, А.Ю.Филиппович. Словник (обратный) Словаря русского языка XI–XVII вв.(вып.1–25). Часть 3 // Словарь русского языка XI–XVII вв. Справочный выпуск. М.: Наука, 2001. С. 393–813.

М.И.Чернышева, Г.Я. Романова, Е.И.Державина. Указатель источников Рукописной древнерусской картотеки (картотеки ДРС) и Словаря русского языка XI–XVII вв. Часть 2 // Словарь русского языка XI–XVII вв. Справочный выпуск. М.: Наука, 2001. С. 265–390.

М.И.Чернышева. Постлексикографический этап: тематические исследования в русской исторической лексикологии. Русская историческая лексикография на современном этапе. К 25-летию издания Словаря русского языка XI–XVII вв. / Отв. ред. Чернышева М.И. М.: ИРЯ РАН, 2000. — С.27–31. — Серия: Отечественная лексикография. Вып. 4.

М.И.Чернышева. Состав и структура Словаря Академии Российской. Словарь Академии Российской 1789–1794. Том 2. М.: МГИ им. Е.Р. Дашковой, 2002. С.12–46.

М.И.Чернышева, Ю.Н.Филиппович. Историко-лексикологическое (тематическое) исследование: экспериментальный опыт на основе информационной технологии. Вопросы языкознания,1999, № 1. С.56–83.

М.И. Чернышева, Г.А Черкасова.. Изменения в тексте переиздания / Словарь Академии Российской 1789-1794 гг. М., МГИ им. Е.Р. Дашковой, 2002-2005 гг. Т.2. С.727-740; Т.3. С.811-828; Т.4. С.760-782; Т.5. С.689-702.

А.Ю. Филиппович. Лингвистический редактор Andrew Tools 2000 / Scripta linguisticae applicatae. Проблемы прикладной лингвистики —

2001. Сборник статей. М.: «Азбуковник», 2001. С.305-310.

А.Ю. Филиппович. Шрифтовое обеспечение электронной версии Словаря Академии Российской 1789-1794 гг. Проблемы построения и эксплуатации систем обработки информации и управления. Сборник статей. Выпуск 7 / Под ред. В.М. Черненького. М.: Кафедра АСОИУ МГТУ им. Н.Э. Баумана, 2005.– С.167–172.

А.Ю.Филиппович Автоматизированная технология корректуры переиздания Словаря Академии Российской 1789-1794 гг. на основе динамически пополняемого словаря спеллера. Вестник Московского государственного университета печати, №5 май. – М.: Изд-во МГУП, 2005 г. – С. 67-85.

Филиппович Анна. Исследование эффективности систем оптического распознавания текстов. // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 7 / Сост. и ред. Ю.Н. Филипповича. – М. Изд-во ООО «Эликс+» 2005. – С. 272-297.

А.Ю. Филиппович. Электронная версия Словаря Академии Российской 1789-1794 годов. Роль книгоиздания в развитии международных научных и культурных контактов: Материалы международной научной конференции (Москва, 21-23 сентября 2005 г.) / Сост. В.И.Васильев, М.А. Ермолаева, А.Ю. Самарин. М.: Наука, 2005. С. 293–296.

Способы представления результатов выполнения проекта

Сведения о ходе работ по проекту и промежуточные результаты разработок предполагается публиковать в форме научных статей и монографий, а также размещать по следующим адресам сети Интернет:

http://www.philippovich.ru — научно-образовательный кластер «Компьютерная лингвистика–искусственный интеллект–мультимедиа» (Computer Linguistics–Artificial Intelligence–Multimedia — CLAIM).

http://www.slovari.ru/ — ООО «Издательский центр «Азбуковник»;

http://www.dashkova.ru/ — Московский гуманитарный институт им.Е.Р.Дашковой (издатель переиздания САР).

На сайтах организаций основных мест работы участников проекта:

http://iu5.bmstu.ru/ — кафедра «Системы обработки информации и управления» МГТУ им.Н.Э.Баумана;

http://www.iling-ran.ru/ — Института языкознания РАН;

http://www.ruslang.ru/ — Институт русского языка им. В.В. Виноградова РАН.

Дополнительные возможности реализованного проекта

Разработанный в проекте CD (DVD) диск электронного издания САР может быть выпущен необходимым тиражом для распространения среди заинтересованных пользователей.

Исследовательские материалы проекта (научный комментарий) и основные компоненты словников, биобиблиографический словарь, указатель источников САР могут быть подготовлены и выпущены в форме печатного научного издания — «Справочного тома» (дополнительного седьмого тома переиздания).