МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Инструментальная среда научных исследований на основе Словаря русского языка XI–XVII вв.

**Ю. Н. Филиппович, М. И. Чернышева,
А. Е. Дубашов, Л. Н. Чуприна**

Инфо-когнитивные технологии, историческая лексикография, словарь русского языка XI–XVII вв., автоматизированные системы научных исследований, электронные издания, автоматическая обработка текстов

A Research Environment Based on the Dictionary of Russian Language of the 11th to 17th Centuries

**Yuriy Philippovich, Margarita Chernysheva,
Aleksey Dubashov, Lyubov' Chuprina**

The report presents the project to create the "Research Environment for Historical and Lexicographical Studies Based on the the Dictionary of Russian Language of the 11th to 17th Centuries. The environment consists of two components: the HiLex system for supporting historical and lexicographical research and an PDF edition of the Dictionary that includes effective search tools for local and remote access.

Словарь русского языка XI–XVII вв. (далее — Словарь или СЛРЯ) представляет собой фундаментальный лексикографический труд, предназначенный как для специалистов, изучающих памятники русской письменности семи веков, а также аспирантов и студентов-гуманитариев, так и для широкого круга читателей. Объем источников и глубина проработки материала позволяет говорить о том, что Словарь, первоначально имевший облик научно-популярного справочника, стал подлинно академическим лексикографическим произведением. Основой Словаря является Рукописная древнерусская картотека (Картотека древнерусского словаря — КДРС). Постоянно пополняемый новыми изданиями Указатель источников к СЛРЯ насчитывает в настоящий момент около 4 тыс. названий, некоторые из которых являются обозначениями собраний в несколько сотен документов, как, например, "Великие Минеи четьи" митрополита Макария [Филиппович Ю., Филиппович А. 2002]. В Словаре представлены все типы источников, это русские летописи, житийная литература (более 120), сказания и повести (около 160), поучения и "Слова" (около 60), писцовые и переписные книги (более 50), приходо-расходные книги (более 80), тысячи грамот и актов, посланий, частных писем ("Грамоток"), предшественники газет — "Вести-Куранты", художественные произведения и проповеди, богослужебные книги и научные труды, юридические памятники, дневники путешественников, отчеты и донесения послов и др. Среди источников есть не только труднодоступные, ставшие библиографиче-

ской редкостью издания памятников и многочисленные рукописи, разбросанные по архивам разных городов, но также плохо сохранившиеся и даже утраченные, которые существуют только в виде карточек. Особенностью источниковой базы Словаря является большое число произведений, переведенных преимущественно с греческого, а также с латинского, польского, немецкого и других языков. Результатом отказа от первоначального лаконизма словарной статьи, когда на значение приводилось только две цитаты, стало снятие ограничения на цитирование, так что каждое значение в Словаре документируется большим числом датированных цитат, а семантика переводных памятников подкрепляется иноязычными параллелями; например, в статье слова *творити* выделено 19 значений, их иллюстрируют 126 цитат, приводится около 80 параллелей.

В настоящее время для обеспечения научных исследований в области русской исторической лексикографии и лексикологии реализуется проект создания Инструментальной среды историко-лексикографических исследований Словаря русского языка XI–XVII вв. (ИСИЛИ). Её использование позволит, во-первых, существенно повысить эффективность работ по исправлению, уточнению и дополнению 1–21 печатных выпусков СЛРЯ новыми материалами; во-вторых, проводить современные лексикологические, лексикографические и тематические исследования материалов Рукописной древнерусской картотеки (КДРС), изданных 22–29 печатных выпусков СЛРЯ, а также новых лексических фактов [Филиппович Ю., Чернышева 2006].

Исходными данными для создания ИСИЛИ являются: 1) материалы СЛРЯ (выпуски 1–21) — сканированные изображения страниц; распознанный текст словарных статей, словники и постраничные индексы заголовочных слов и словоформ цитатного материала; 2) база данных электронного указателя источников КДРС и СЛРЯ; 3) словник (список слов-разделителей) КДРС; 4) тексты СЛРЯ (выпуски 22–29).

ИСИЛИ состоит из двух комплексов: системы поддержки историко-лексикографических исследований HiLex (адаптированная версия системы создания и ведения лексикографических картотек WinDialex [Дубашов 2006] в комплексе с другими программными средствами) и электронного издания СЛРЯ.

WinDialex — это комплекс программных средств, предназначенный для автоматизированной обработки корпусов текстов с целью создания и ведения таких лексикографических объектов, как словники (простые и частотные), конкордансы (полные и локальные), словоуказатели, терминологические указатели и т. п.

HiLex (HistoryLexicographicalsystem) состоит из следующих компонентов: 1) системы WinDialex, адаптированной к целям и исследовательским задачам проекта, которая допускает использование в качестве исходных данных структурированные словарные тексты; 2) комплекса программных средств Flotatation (“Флотация”), необходимого для формирования макетов новых статей СЛРЯ на основе уже существующих, построения словарных проекций на определенную

букву; 3) лексикографической библиотеки, предназначенной для разработки лексикографических программных продуктов.

Предполагается одновременное существование двух версий электронного издания СЛРЯ: а) исходного (идентичного печатному изданию 1–21 выпусков) — статичного, неизменяемого; б) текущего (содержащего все поправки, исправления, дополнения исходного издания и 22–29 выпуски), постоянно пополняемого коллективом авторов проекта. Обе версии издания будут снабжены эффективным справочно-поисковым аппаратом, позволяющим осуществлять символьный поиск словарных статей, словоформ цитатного материала и их синтагматических конструкций. В качестве формата электронного издания выбирается PDF. В структуре электронного издания можно выделить подсистемы "настольного" и веб-издания для локального и удаленного доступа к ресурсам СЛРЯ.

Настольное издание реализуется в виде комплекта файлов формата pdf. Каждый файл комплекта соответствует одному из томов Словаря. Файлы снабжены закладками для навигации между статьями Словаря и постраничными комментариями. Комментарии представляют собой текст страницы Словаря. Наличие комментариев позволяет выполнять поисковые операции, т. е. поиск по словам и сочетаниям слов, встречающимся в цитатном материале. Сформированное издание может открываться с помощью любого средства для просмотра файлов формата pdf. Для автоматизации создания pdf-издания Словаря разработан специальный программный модуль формирования закладок и комментариев к страницам в готовом файле формата pdf.

Программный комплекс электронного издания СЛРЯ ориентирован на использование стандартного пользовательского интерфейса и исполнения на персональном компьютере пользователя. В качестве архитектуры выбрана клиент-серверная технология. В качестве средства для разработки модуля ввода данных и модуля для создания pdf-издания будет использован язык Java и среда разработки NetBeans версии 6.9.1.

В качестве СУБД выбрана MySQL. Эта СУБД является свободной, распространяется под GeneralPublicLicense или под собственной коммерческой лицензией. MySQL является решением для малых и средних приложений. MySQL имеет API для языков Delphi, C, C++, Эйфель, Java, Лисп, Perl, PHP, Python, Ruby, Smalltalk и Tcl, библиотеки для языков платформы .NET, а также обеспечивает поддержку для ODBC посредством ODBC-драйвера MyODBC, что позволяет легко организовать работу модуля ввода данных.

Веб-интерфейс пользователя реализован с помощью языка PHP 5. Для обработки http-запросов, получаемых от пользователя, и генерации html страниц для отправки пользователю будет использован Apache HTTP-сервер. Apache является кроссплатформенным ПО, поддерживает операционные системы Linux, BSD, Mac OS, MicrosoftWindows, NovellNetWare, BeOS. Для разработки и отладки веб-приложения используется набор дистрибутивов "Denwer".

При создании электронного издания СЛРЯ будут использоваться новые методы, которые касаются и теоретического, и экспериментального подходов. В теоретическом плане это инновационная методика нечеткого индексирования графических форм текстовых данных на основе их распознанных копий. Экспериментально данная методика реализуется на сверхбольшой базе текстовых данных, составляющей более 500 тыс. различных словоформ, что позволяет обеспечить высокий уровень релевантности поисковых операций.

Литература

Дубашов 2006 — *Дубашов А. Е.* Методы и алгоритмы извлечения данных из словарных текстов: на примере Словаря русского языка XI–XVII вв.: дис. ... канд. техн. наук: 05.13.06. Москва, 2006.

СЛРЯ — *Словарь русского языка XI–XVII вв*. Вып.1–29. М.: Наука, 1975–2011.

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

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

Чернышева 2006 — Словарь русского языка XI–XVII вв.: дополнения и исправления: тетрадь первая: А–Б. М.: Наука, 2006. 61 с.