УДК 81'33
UDC 81'33

Бурыкин Алексей Алексеевич
Институт лингвистических исследований
Российская Академия Наук
г. Санкт-Петербург, Российская Федерация
Alexis A. Burykin
Institute for Linguistic Studies
Russian Academy of Sciences
St. Petersburg, Russian Federation
e-mail: albury@rambler.ru

ЭЛЕКТРОННЫЙ РЕСУРС ДЛЯ ИССЛЕДОВАНИЙ В ОБЛАСТИ РУССКОЙ ЛЕКСИКОЛОГИИ И ЛЕКСИКОГРАФИИ «БИБЛИОТЕКА ЛЕКСИКОГРАФА»: ОПЫТ РАБОТЫ, ПЕРСПЕКТИВЫ ПОПОЛНЕНИЯ, ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ
ELECTRONIC RESOURCE FOR STUDIES IN THE FIELD OF RUSSIAN LEXICOLOGY AND LEXICOGRAPHY «LEXICOGRAPHER'S LIBRARY»: EXPERIENCE, OUTLOOK FOR ENLARGING, POSSIBILITIES OF USING

Аннотация
Настоящая статья посвящена использованию информационных технологий для обработки филологического материала. Даётся обзор доступных электронных библиотек, достоинств и недостатков организации в них текстового материала, подчёркивается необходимость дальнейшей оптимизации работы в области русской лексикологии и лексикографии. Подробно анализируется созданный корпус «Библиотека лексикографа». В Библиотеке насчитывается 3,1 млрд. словоформ из более чем 260 тыс. текстов различных видов. Максимальная полнота охвата материала, открытость и возможности для неограниченного пополнения и редактирования являются достоинствами проекта. Объективным недостатком проекта является его принципиально оффлайновый характер. Рассматривается проблематика приведения текстового материала к единому формату и повышения эффективности поисковых систем, в том числе в связи с вариантностью графико-орфографической репрезентации материала в разные хронологические периоды. Приводится спектр задач, которые можно решать с помощью «Библиотеки лексикографа», включая описание, уточнение, хронологизацию, систематизацию, восполнение материала.

Abstract
The current paper deals with applying information technology for philological material processing. The review of available electronic libraries is given, advantages and disadvantages of arranging texts within those libraries are viewed, the necessity of further optimization of work in the field of Russian lexicology and lexicography is stressed. Detailed analysis of the project «Lexicographer's Library» is presented. The Library holds 3,1 billion

word forms from more than 260 thousand texts of various types. Maximal coverage of the material, open status and unlimited possibilities of enlarging and editing the materials are the advantages of the project. Its natural disadvantage is its intentionally offline status. The problematic issues of format unification and increasing the effectiveness of search systems particularly connected with variance of graphic representation and spelling in different time periods of development of the system of writing are addressed. The range of tasks that can be performed using «Lexicographer's Library» is demonstrated including description, specification, chronological arrangement, systematization, restoring lost or missing segments.

Ключевые слова: прикладная лингвистика, компьютерная лингвистика, электронные ресурсы, корпус текстов, лексикология, лексикография, картотека, тезаурус.

Keywords: applied linguistics, computational linguistics, electronic resources, corpus, lexicology, lexicography, card store, thesaurus.

Компьютеризация филологических и конкретно – лингвистических исследований, однозначно состоявшаяся в 1990-е годы, потребовала и продолжает требовать такой формы исходного языкового материала, которая была бы не только пригодна и доступна, но и удобна для работы в любых условиях и для любого пользователя. Одним из критериев удобства использования в данном случае, как мы полагаем, окажется доступность материала, то есть возможность его использования в каких угодно условиях начиная от Интернета или компьютерной сети крупного научного центра до персонального компьютера студента-первокурсника.

Использование новых информационных технологий для обработки филологического материала привлекло внимание лингвистов самых разных направлений, и, в первую очередь, специалистов по прикладной лингвистике, в недрах которой оформилась самостоятельная дисциплина – корпусная лингвистика. В настоящее время корпусная лингвистика стала предметом описания в специальных статьях и учебных пособиях [Зубов 2004; Захаров 2005; Перцов 2006; Баранов 2007; Белозерова, Чуфистова 2007; Тетакаева 2011; Захаров, Богданова 2011; Грудева 2012; Богданова 2012], проблемы корпусной лингвистики активно обсуждаются на специальных конференциях (http://www.dialog-21.ru/digest/, http://corpora.phil.spbu.ru (и вкладки на этом сайте), см. также http://corpora.iling.spb.ru, http://corpora.iling.spb.ru/theory.htm). Значительным достижением отечественной корпусной лингвистики справедливо считается Национальный корпус русского языка [Национальный корпус ...: 2003–2005].

Литература по корпусной лингвистике и отдельным аспектам применения корпусов становится уже труднообозримой. В её рамках те теоретические и практические проблемы, которые выглядят вечными проблемами традиционной филологии, а именно – лексикология и лексикография, – както теряются в колоссальном объёме отдельных частностей типов корпусов, их устройствах, специальных подразделениях корпусов и так далее. Между тем, практическая работа над словарями русского языка, в том числе над «Большим академическим словарём современного русского языка», а также

необходимость критической оценки ранее подготовленных словарей русского языка, словарей русского языка, появляющихся на отечественном книжном рынке, заставляет задумываться над качественным изменением ресурсов для словарной работы. В равной мере сказанное относится к исследованиям по русской лексикологии, будь то описание семантики и истории отдельных слов, хронологического пласта лексики, тематической или лексико-семантической группы лексики. В этой сфере также отчётливо обозначается две задачи: во-первых, дать в распоряжение исследователям ресурсы современного уровня, призванные заменить книжные издания текстов и бумажные картотеки (см. [Рогожникова 2003]), во-вторых, разработать инструмент для верификации выполненных и выполняемых лексикологических исследований. Что касается электронных картотек – это особая и весьма проблемная область прикладной лингвистики, в которой и имеется немало проблем [Захаров 2007]. Вообще, будущее электронных картотек, как нам кажется, непонятно и туманно. Этот продукт прикладной лингвистики будет страдать теми же недостатками, что и бумажные картотеки – ограниченностью объёма источников и субъективностью их подбора, навязываемыми пользователю границами цитаты, прецедентностью толкований (если таковые есть), ограниченными возможностями верификации хронологии появления лексических единиц в языке и тому подобными.

В настоящее время, а именно – в последние 10–15 лет, текстовые ресурсы для работы над корпусами и текстовыми базами данных не являются дефицитом. Наиболее известным и широко доступным источником русских текстов являются электронные библиотеки, доступные в Интернете (Библиотека Мошкова, Альдебаран, и т. д.), те же библиотеки, выпущенные на компакт-дисках (Библиотека «Всемирная литература», та же Библиотека Мошкова с частью ресурсов), сайты, посвящённые творчеству отдельных писателей, электронные собрания сочинений отдельных писателей XIX–XX веков, тематические собрания текстов (так, были выпущены диски с образцами русской драматургии и русской поэзии XIX-XX веков). Однако все эти продукты, несмотря на их полезность и такое немаловажное достоинство, как высокое качество текстов, имеют ряд недостатков технического порядка. Отдельные тексты, извлекаемые из электронных изданий, с большим трудом поддаются объединению в единое собрание, поскольку для этого требуется их дополнительная обработка. Электронные библиотеки, как в ресурсах Интернета, так и на компакт-дисках, сильно разнятся по своему составу, причём объём библиотеки не определяет её качества как электронного ресурса. В таких собраниях переводная художественная литература заметно доминирует количественно над отечественной литературой, а в последней – современные боевики и детективы преобладают над классикой. Единая для диска поисковая система (такая, какая имеется, к примеру, в Библиотеке «Всемирная литература») при применении её к лексическому материалу собрания текстов хотя и действует, но оказывается бесполезной, она не предоставляет возможности ни отобрать нужный фразовый или текстовый материал, ни извлечь его.

Необходимость дальнейшей оптимизации работы в области русской лексикологии и лексикографии, причём как синхронной, ориентированной на тексты и словарный состав современного русского литературного языка, так и исторической (в пределах конца XVII–XX веков) побудила автора настоящей работы приступить к реализации проекта электронной библиотеки русских текстов, которая была бы предназначена специально для филологических исследований и была бы адресована пользователям-филологам.

Данный проект, получивший название «Библиотека лексикографа», реализуется автором в Словарном отделе ИЛИ РАН с начала 2008 года [Бурыкин 2008, 2010, 2013] по настоящее время и будет продолжаться. Источником для собирания материала являются все доступные в Интернете электронные библиотеки и тематические сайты (в настоящее время обследовано около 1300 ресурсов), а также собрания текстов на компакт-дисках, сыгравшие определённую роль на начальном этапе работы над проектом.

Материал для «Библиотеки лексикографа» (далее – Библиотека) отбирается примерно по тем же принципам, по которым комплектовалась и продолжает комплектоваться Библиотека Словарного отдела ИЛИ РАН, которая в течение многих лет является хранилищем источников для Большой словарной картотеки ИЛИ РАН. Хронология материала в описываемом проекте охватывает период от начала XVIII века до начала XXI века. Автором проекта ведётся собирание древнерусских текстов XI–XVII вв. в электронном виде, но этот материал имеет ограниченное применение по причине разнообразных различий в техническом оформлении оригинального древнерусского текста и передаче графики текста. В соответствии с этим в собрание текстов «Библиотека лексикографа» включаются следующие виды электронных документов:

- художественная литература;
- литературно-художественная критика, публицистика;
- общественно-политическая литература;
- мемуары политических деятелей, деятелей науки, культуры, искусства, военные мемуары;
- материалы из периодики;
- официальные документы, законодательные акты и тому подобное;
- научно-популярная литература и учебные пособия по всем областям знаний.

Какими же достоинствами обладает проект «Библиотека лексикографа» по сравнению с Национальным корпусом русского языка и иными ресурсами русских текстов и лексического материала русского языка?

Прежде всего, это объём ресурса – в «Библиотеке лексикографа» насчитывается 3,1 млрд. словоформ из более чем 260 тыс. текстов (от отдельного стихотворения и газетной заметки до объёмных романов).

Далее, это независимость от подключения к Интернету – отнюдь не везде и далеко не все исследователи имеют возможность часами эксплуатировать Интернет в поисках чьих-либо мемуаров и в течение продолжительного времени работать с Национальным корпусом русского языка. Далее, при составлении «Библиотеки лексикографа» автор проекта ориентируется

на максимальную полноту охвата материала по указанным разделам, и эта задача при изучении десятков и сотен электронных библиотек имеет относительно успешное решение. В сфере художественной литературы тексты писателей русского зарубежья, ставшие доступными как в печатном, так и в электронном виде за последние 30–35 лет – таких, как Б. Зайцев, И. Шмелев, М. Осоргин, Г. Газданов, Н. Берберова, и так далее, а также произведения опальных отечественных писателей (Е. Замятин, Б. Пильняк, А. Веселый и др). Тексты сочинений политических и военных деятелей (Л. Троцкий, П. Врангель, А. Деникин. П. Краснов и т. д.) восполняют лакуну в отечественных лексикографических ресурсах, поскольку сочинения этих авторов ранее никогда не привлекались для пополнения словарных картотек. В рамках нашего проекта имеется возможность разместить в нём в максимально полном виде тексты таких авторов, как М. Булгаков, М. Зощенко, М. Цветаева, А. Ахматова, О. Мандельштам, Б. Пастернак, Б. Окуджава и многих других, чьи тексты в лексикографической работе почти не использовались, или, по крайней мере, их воспроизведение в словарях не приветствовалось. Автором проекта ставится задача не только подготовить качественно новый продукт, предназначенный для лексикографических работ и лексикологических исследований, но и за счёт доступных ресурсов – в идеале или в отдалённой перспективе – воссоздать в электронном виде корпус источников имеющихся словарей и, соответственно, – весь материал Большой словарной картотеки ИЛИ РАН, в том числе и ту его часть, которая не была помечена при разметке источников для пополнения картотеки.

Достоинствами данного проекта являются его открытость и возможность для индивидуального пополнения всеми пользователями, а также возможность редактирования проекта посредством перегруппировки текстов и создания собственных собраний текстов для решения тех или иных задач. «Библиотека лексикографа» по замыслу и по условиям применения – открытый ресурс, в котором пользователь может добавлять в него новые тексты или удалять или временно выводить из оборота ненужные ему тексты (напр., более ранние или более поздние, тексты определённых жанров или тематики), а также заменять одни электронные версии текстов новыми, более качественными и более авторитетными. Закрытые корпуса текстов не позволяют этого делать, а отсеивание неиспользуемых материалов картотеки может производиться только вручную при работе с каждой отдельно взятой словарной единицей. Обновление картотеки при наличии выборки из текстов одного и того же автора даже в небольшом объёме (3–4 тыс. примеров), например, при появлении более совершенного «академического» издания сочинений какого-либо автора, по существу невозможно.

Нетривиальным преимуществом данного проекта является то, что его пользователи не связаны с разрешительным доступом к картотеке и не ограничены в своей работе с проектом какими-то часами работы структурных подразделений ИЛИ РАН – они могут работать с ним индивидуально вне стен Института.

Объективным недостатком проекта для тех, кто знаком с ним, является его принципиально оффлайновый характер и невозможность пользоваться

им при посредстве Интернета. Для подобной позиции у создателей проекта есть свои основания: прежде всего – это довольно жёсткая конкуренция на рынке словарных изданий по русскому языку, это определённые амбиции отдельных лексикографических коллективов в продолжении работы над объёмными словарями русского языка, также придающие работе конкурентный характер, наконец, это некоторая внутренняя конкуренция «картотека против корпуса» в стенах самого Института лингвистических исследований, которая была связана со сменой поколений в Словарном отделе ИЛИ РАН и созданием в Институте иных подразделений со сходными задачами – Лаборатории автоматизации лингвистических исследований (1986–1997), Лаборатории Информационных лингвистических технологий и так далее. Технический недостаток «Библиотеки лексикографа» – отсутствие какой-либо разметки, хотя при предназначенности корпуса исключительно для лексикологии и лексикографии разметка общего характера не является необходимой, а специальная (снятие омонимии, семантическая разметка) – весьма затруднительной для выполнения.

Во второй половине 2000-х годов специалистами по лексикографическим ресурсам в ИЛИ РАН неоднократно поднимался вопрос о качестве электронных текстов и степени их достоверности для разработок в области лексикологии и в особенности для академической лексикографии, предусматривающей в рамках издательского цикла сверку цитат с материалами картотеки. С этой целью было принято решение размещать в Библиотеке произведения, представляющие русскую художественную литературу, в нескольких вариантах – отдельные произведения (при этом такие романы, как «Война и мир» Л. Толстого, или «Тихий Дон» М. Шолохова размещаются и в виде одного файла с полным текстом, и в виде нескольких файлов по отдельным частям) и собрания сочинений по томам: идентичность фрагментов текста в потенциальных цитатах является свидетельством достоверности цитаты и точности передачи текста. Впрочем, текстологические ошибки, опечатки и лексические аберрации, искажающие текст, но не отслеживаемые ни автоматическими корректорами, ни чтением текста, могут присутствовать в любой словарной картотеке. Практика работы со словарными картотеками показывает, что там не всегда учитываются более авторитетные повторные издания текстов, не принимаются во внимание разночтения канонических текстов с первыми изданиями, не ведётся мониторинг текстологических уточнений по позднейшим изданиям текстов или исследованиям соответствующей рубрики «Заметки. Уточнения» в журнале «Русская литература».

Художественная переоценка ценностей и устранение идеологических рамок в области русской литературы XX века приводит к тому, что произведения некоторых авторов, составлявшие основу для первого издания 17-томного словаря современного русского языка – например, романы С. Бабаевского «Кавалер Золотой звезды». М. Бубеннова «Белая берёза» и тому подобные – разыскиваются в электронной форме с большим трудом, но, поскольку они присутствуют на сайтах «патриотической» ориентации и появляются в библиотеках с расширением их объёма, то они все же занимают своё место в

«Библиотеке лексикографа». В ней уже присутствуют – рядом с книгами современных политиков, представляющих самые разнообразные течения и партии, – и «Краткий курс истории ВКП (б)», и работы В. И. Ленина в том количестве, в каком их удаётся отыскать (доступно 5-е издание собрания сочинений в 55 томах), и сочинения И. В. Сталина (всё 16-томное собрание сочинений), сочинения Н. С. Хрущева, изданные до 1964 года, и отдельные сочинения Л. И. Брежнева.

Кстати, выбор текстов из электронных библиотек для пополнения ресурсов «Библиотеки лексикографа» требует основательного знания «советской» литературы 1930–1980 годов, поскольку приоритетными для проекта по ряду соображений являются тексты таких авторов, как Л. Леонов, К. Паустовский, В. Каверин, В. Катаев, Д. Гранин, В. Шефнер, В. Белов, В. Астафьев, В. Липатов, А. Лиханов, Е. Носов и другие – те, чьё творчество представляет русский язык середины и второй половины XX века. Материал современной литературы (Н. Леонов, М. Веллер, А. Маринина, Д. Донцова, Л. Улицкая, Е. Вильмонт, М. Серова и т. п.) пока включается в Библиотеку выборочно, отдельными образцами. Довольно досадно, что произведения некоторых авторов по существу недоступны в электронных версиях – так, из произведений Вс. Кочетова в Интернет-библиотеках представлен только роман «Чего же ты хочешь?», в то время как другие романы этого автора, сохраняющие свою ценность (напр., «Угол падения»), в электронном виде отсутствуют. Из сочинений В. Ажаева во множестве библиотек имеется роман «Вагон» в то время как более ранние и отнюдь не менее известные произведения этого автора (к примеру, роман «Далеко от Москвы») доступны с большим трудом. Не без труда отыскиваются ранние произведения В. Аксенова («Коллеги»), Д. Гранина («Искатели», «Иду на грозу»). Поиск сочинений отдельных авторов занимает длительное время, в основном сопряжённое с ожиданием электронных версий в тех или иных доступных собраниях.

Обработка электронных материалов, извлекаемых из библиотек Интернета, предполагала и предполагает приведение их к единому формату, который облегчал бы и работу поисковых программ, и извлечение фрагментов текстов для работы. В качестве единого рабочего формата для помещаемых в Библиотеке текстов избран формат TXT. Для такого выбора было довольно много причин: компактность файлов с текстами, наличие большого числа файлов в этом формате во многих библиотеках, удобство конвертации в данный формат текстов, копирующихся с сайтов, где они выставлены в незаархивированной форме, возможность прямой конвертации файлов, извлекаемых из электронных библиотек, содержащих книги для чтения на карманных компьютерах в формате FB2 и ему подобных. Для пользования Библиотекой нами применяется текстовый редактор Bred (версии 2 или 3), позволяющий работать с txt-файлами любого объёма и решать основную массу проблем с кодировками текста. При технической обработке текстов, в частности, при их конвертации из форматов DOC и RTF в дополнение к текстовому редактору Word используются текстовые редакторы Atlantis, Hieroglyph, Polyedit, Ultraedit и так далее, которые

иногда лучше работают с файлами большого объёма и опции которых могут быть использованы при очистке электронной формы текста от лишних элементов (служебные пометы, реклама, ошибки сканирования и распознавания и т. п.). Единственным ограничением на использование тех или иных ‘электронных копий источников является качество файлов формата PDF или Djvu, не всегда позволяющее получить качественный файл в редактируемом формате (к досаде, не удаётся пока разместить в Библиотеке тома из собрания сочинений Г. Е. Зиновьева, издававшееся в 1920-е годы).

Важной составляющей «Библиотеки лексикографа» являются поисковые программы. Несомненно, самой удобной поисковой программой для «Библиотеки лексикографа» оказывается программа Archivarius3000 в любой доступной версии. Эта программа позволяет просматривать все словоупотребления запрашиваемого слова во всех текстах, помещённых в Библиотеке, причём без ограничения количества просматриваемых цитат, она открывает доступ к каждому конкретному тексту с возможностью скопировать из текста цитату любого объёма, и она также позволяет формировать и копировать пакеты текстов, в которых встречается искомая лексическая единица. Важным достоинством этой программы является возможность её целенаправленного настраивания на данный ресурс так, чтобы при поиске материала ей не захватываются все иные материалы, присутствующие в компьютере. Повторное индексирование ресурса осуществляется по мере пополнения «Библиотеки лексикографа» новыми материалами и не отнимает много времени, хотя требует наличия свободного места на системном диске или иных жёстких дисках компьютера.

Из технических требований к компьютерам для работы с электронными ресурсами, в частности, с отобранными текстами и с самой «Библиотекой лексикографа», наиболее значимыми являются не столько объём жёсткого диска, на котором размещается ресурс, сколько быстродействие процессора и объём оперативной памяти, а также скорость работы самого жёсткого диска.

В дальнейшем предполагается, кроме обобщающего корпуса текстов, выделить в нём отдельные модули – тексты XVIII, XIX и XX веков с внутренним делением по жанрам (поэзия, литературная проза, нехудожественная проза, документы, общественно-политическая литература, научно-популярная литература, историческая литература (труды историков и исторические романы), юмор и сатира. Целесообразно выделить в отдельный модуль произведения современной художественной литературы (без оценки жанров и достоинств отдельных авторов) специально для отслеживания использования новых слов в русском языке. Ещё раз отметим, что важным достоинством «Библиотеки лексикографа» является и то, что она легко может быть откорректирована, настроена и пополнена в соответствии с индивидуальными или коллективными запросами любых пользователей – тех, кто занимается исследованиями лексики русского языка, специалистов по исторической лексикологии, тех, кто изучает лексикостилистику и язык писателей, наконец, данное собрание может быть «настроено» для составления толкового словаря русского языка любого типа, объёма и хронологического диапазона.

Проект «Библиотека лексикографа» задумывался и поначалу действовал как электронная альтернатива Библиотеке словарного отдела ИЛИ РАН, которая служила и служит источником бумажной Большой словарной картотеки. В настоящее время, при насыщении Библиотеки соответствующим объёмом текстов и наличии программного обеспечения – в ней используется поисковая программа Архивариус3000, – функции электронного собрания расширились и могут быть охарактеризованы следующим образом.

1. Собственно **электронная библиотека**, позволяющая обращаться к любому входящему в неё тексту с любыми целями.

2. **Виртуальный тезаурус лексики русского языка**, позволяющий осуществлять выбор текстов, содержащих ту или иную лексическую единицу, и все примеры её употребления. Эта возможность, реализующаяся при необходимости сохранения документации на отдельные конкретные слова, может быть осуществлена при использовании любой поисковой системы, начиная с опции поиска в операционной системе Windows, или поисковых программ наподобие программы Integra, позволяющей просматривать цитаты из текстов.

3. **Виртуальная электронная картотека**, включающая все лексические единицы всех текстов, присутствующих в Библиотеке. В настоящее время общий массив лексики, к которой обеспечивается доступ в виртуальной картотеке, составляет около 3,1 млрд. словоформ, из которых более 7 млн. единиц составляют разные слова (статистические данные содержатся в отчётах программы Архивариус3000 при индексировании текстов Библиотеки). Поисковая программа Архивариус3000 позволяет не только просматривать все употребления любого заданного слова в текстах, но и указывает количество употреблений данного слова в каждом тексте и даёт возможность увидеть их; она же открывает возможность самостоятельного выбора необходимого фрагмента текста, документирующего то или иное слово, и копирования такого фрагмента для размещения в тексте составляемого словаря или исследования.

Нарастание объёма текстов, размещаемых в «Библиотеке лексикографа», а также пользование ей при работе над словарями, охватывающими разные периоды истории русского литературного языка, ставит перед создателем проекта и его пользователями новые задачи (см. [Герд, Захаров, 2004; Герд, 2013]), к числу которых относятся хронологическая классификация текстов и распределение текстов по определённым тематическим рубрикам и жанровым и стилистическим формам.

Средством решения этих задач, предусмотренным в рамках модернизации «Библиотеки лексикографа» и оптимизации её использования, является присвоение определённых условных индексов всем документам-файлам Библиотеки, которые включаются в состав имён файлов. Имя файла содержит фамилию и инициалы (или первый инициал) автора и название произведения или собрания произведений (стихи, повести и рассказы, романы и т. п.). К этой информации добавляются сведения о хронологической отнесённости документа, его жанровой форме и предметно-тематической отнесённости. Для этого автором проекта предложено использование букв латинского алфавита, которые отсутствуют в именах файлов или могут быть устранены оттуда при их наличии.

Относительная датировка документа предполагает его привязку к периоду в границах одной трети XVIII, XIX, XX и XXI веков. В качестве кодовых символов для этих периодов использованы четыре последние буквы латинского алфавита в следующих комбинациях:

WW – первая треть 18 века,
WWW – вторая треть 18 века,
WWX – последняя треть 18 века,
XXW – первая треть 19 века,
XXX – вторая треть 19 века,
XXY – последняя треть 19 века,
XYY – первая треть 20 века,
YYY – вторая треть 20 века,
YYZ – третья треть 20 века,
YZZ – первая треть 21 века,
ZZZ – вторая треть 21 века.

Количественная осложнённость символов позволяет осуществлять автоматически выборку текстов, относящихся к нескольким смежным периодам истории русского языка, и задавать самую различную структуру запросов. Так, поиск по символу W даст тексты, относящиеся к 18 и первой трети 19 века, поиск по символу X – тексты, относящиеся к последней трети 18 века, всему 19 веку и первой трети 20 века, поиск по символам XX позволит сгруппировать тексты, относящиеся к 19 веку, а поиск по символам XY – выделить тексты конца 19 – начала 20 веков. В этой систематике учтена реальная периодизация истории русского литературного языка нового времени, отражающаяся на изменениях в его словарном составе. Эти индексы позволят автоматически выбирать из объёма Библиотеки тексты, укладывающиеся по времени создания в периоды, равные трети века, в хронологическом интервале Библиотеки – XVIII – начало XXI вв. в объёмах, равных 30, 60, 100 и 130 годам (треть или две трети любого века, век целиком или век с третью предшествующего или последующего века, смежные 60-летние периоды двух веков), а также тексты любой жанровой формы и любой тематики. Хронологические и, по необходимости, жанровые или тематические выборки существенно уменьшают объём просматриваемого материала при работе с высокочастотными лексическими единицами, в то же время присутствует возможность обращаться к смежной по времени или иной по жанру выборке текстов.

Жанрово-стилистическая характеристика текстов представлена в Библиотеке в следующей классификации (перед наименованием рубрики здесь и даже приведён литерный индекс):

a) художественная литература;
b) публицистика, критика;
c) мемуары, дневники, переписка;
d) общественно-политическая литература, научная, научно-популярная литература;
e) документы, официальные материалы (партийные программы, государственные законы, указы, стенограммы, документация, итоговые документы мероприятий, служебная документация (приказы, циркуляры и т. п.);

f) газетно-журнальная периодика;

g) переводная литература и иные переводные источники (резервная позиция).

Предметно-тематическая отнесённость источников в предварительной форме (она открыта для обсуждения) имеет следующий вид:

h) гуманитарные науки (философия, религия и религиоведение, правоведение, политология и социология, демография, экономика, культурная антропология и культурология и искусствоведение, литературоведение и языкознание и книговедение, психология и педагогика);

i) история (исторические труды, исторические романы, биографические произведения);

j) науки о земле и человеке (география, геология, биология, антропология, медицина и ветеринария, секс и сексология);

k) путешествия, география, страноведение, этнография, страны мира, народы мира;

l) бытовые практики (быт и повседневные домашние практики, жилище и его устройство, гигиена, пища и её приготовление, одежда и её изготовление и ремонт, ремесло – производство и ремонт предметов, торговля, досуг, хобби, игры и развлечения, спорт, домашние животные и культурные растения – сад, огород, комнатные цветы);

m) точные и естественные науки (математика, физика, химия, астрономия, навигация и т. д.);

n) производство, техника, строительство (промышленное производство и его организация, промышленная техника и её история, промышленное строительство и т. д.);

o) связь, коммуникационные технологии, технологии хранения и обработки информации (почта, телеграф, телефон, компьютеры и т. д);

p) транспорт (сухопутный транспорт, гужевой, автомобильный, транспорт, водный, морской и воздушный транспорт);

r) военное дело, военное искусство (теория, история, практика военного дела, военные мемуары и биографии, художественная военная литература);

s) морское дело (теория, история, практика морского дела, описания походов, мемуары и биографии, морской и водный транспорт, морская художественная литература);

t) воздухоплавание и авиация, космонавтика (теория и история воздухоплавания и авиации, военная авиация, воздушный транспорт, мемуары и биографии, художественная литература об авиации и космонавтике);

u) переводная научно-техническая литература (резерв).

Отличие предлагаемой классификации от систематики книг в электронных библиотеках и от принятой в библиотековедении универсальной классификации состоит в том, что она является нежёсткой и не только допускает, но и предполагает включение одного и того же текста в несколько разных тематических групп.

Автоматический поиск текстов по индексам, вставленным в имена файлов, что может осуществляться при использовании опции поиска в Windows без использования других программ, позволит осуществлять вы-

бор источников как по хронологическим, так и по жанровым и тематическим характеристикам и по свойствам текстов в любой комбинации. Внутри тематических рубрик при использовании цифр от 1 до 0 может быть использована более детальная тематическая классификация документов.

Количественная оценка объёма источников, представляющих те или иные предметно-тематические рубрики и области знаний и практики, позволит поставить задачи пополнения «Библиотеки лексикографа» за счёт дополнительных поисков электронных текстов соответствующей тематики и сканирование тех текстов, которые представляются необходимыми для использования в лексикологических исследованиях и лексикографической работе.

В настоящее время «Библиотека лексикографа» существует в двух вариантах – в виде единого собрания файлов, сгруппированных по алфавиту имён файлов (имён авторов или названий документов), и в виде собрания датированных текстов, сгруппированных по годам начиная с 1700 года по 2015 год. Оба варианта – как, впрочем, и неограниченное количество вариантов Библиотеки, созданных пользователем по своему желанию – размещаются под одной программной оболочкой внутри одной копии программы Архиваниус3000 с разными индексами. Последний вариант, позволяющий с большой точностью определять время вхождения тех или иных слов в русский язык, а также устанавливать авторство некоторых неологизмов, востребован в практике составления словарей Новых слов, он позволяет группировать по периодам лексику русского языка XVIII и XIX веков при работе над соответствующими словарями, составление которых осуществляется в Словарном отделе ИЛИ РАН.

Возможности проекта «Библиотека лексикографа» позволяют решать следующие задачи:

1) обобщение и систематизация лексики русского языка, характерной для русского языка нового времени или какого-то периода его истории (напр., XVIII, XIX, XX века и т. п.), описание специфических лексических единиц и особенностей семантики известных лексических единиц, характерные для того или иного периода;

2) изучение динамики в семантике или сочетаемости слов;

3) восполнение лакун ранее изданных словарей русского языка, охватывающих по хронологии данный период: прежде всего – фиксация и описание лексических единиц, отсутствующих в изданиях БАС и других словарях современного русского языка;

4) пополнение объёма иллюстративного материала: уточнение описаний значений и стилистических характеристик слов;

5) уточнение хронологии вхождения отдельных слов в обиходный русский язык;

6) изучение аксиологических характеристик лексических единиц и обозначаемых ими понятий.

По существу рассмотрение аксиологических, ценностных свойств лексических единиц и тех понятий, которые они выражают, на основе словарей русского языка выглядит бесперспективным – в плане аксиологии тут за того современного исследователя, который задумался над ценностными состав-

ляющими тех или иных концептов и констант русской культуры, постарался лексикограф, составлявший ту или иную словарную статью 20, 30 или 60 лет назад, а также выборшик цитат для картотеки. Ресурс «Библиотека лексикографа» даёт возможность исследователю реализовать первооткрывательский доступ к большинству иллюстративных цитат, документирующих ту или иную лексическую единицу или репрезентирующих тот или иной концепт.

Изучение«редких и забытых слов», вылившееся в конце прошлого столетия в составление целого ряда словарей (см. [Сомов, 1996; Рогожникова, 1996; Рогожникова, Карская, 1997; Елистратов, 1997; Глинкина, 1998] и др.) ставил перед специалистами по историчесекой лексикологии две задачи – во-первых, продолжение исследований в этой сфере на новой и стандартной единой базе, во-вторых, уточнение данных, полученных ранее «вручную» на бумажных картотеках и при чтении текстов ранее. Здесь следует ожидать как увеличения объёма иллюстративного материала, так и уточнения хронологических данных по времени вхождения отдельных слов в русский литературный язык.

Ресурс «Библиотека лексикографа» открывает возможности поиска и описания не только отдельных лексических единиц, но и словосочетаний, в первую очередь устойчивых словосочетаний (коллокаций), а также фразеологических единиц, пословиц и поговорок, крылатых слов. Он позволяет не только с большой точностью и достоверностью учитывать и представлять варианты фразеологизмов и идиом иных типов, но и описывать своеобразную фразеологическую грамматику – ограничения на грамматическое изменение тех или иных идиом, хотя пока в рамках корпусной лингвистики грамматика представлялась как объект статистических наблюдений [Копотев, 2008]. В корпусных исследованиях уже обращалось внимание на возможности изучения статистических характеристик материала [Соловьев, 2011] и на исследования идиоматики [Плисецкая, 2013]. Значительный объём иллюстраций на фразеологизмы позволяет решать такие задачи, как создание Частотного словаря русской фразеологии с использованием словников доступных фразеологических словарей, а также и всесторонних, в том числе и относительно-статистических характеристик таких единиц, как пословицы и поговорки в письменных текстах, и крылатые слова – тут перспективы обещает сравнение данных по «Библиотеке лексикографа» с имеющимися наблюдениями филологов (см. [Русская судьба..., 2010]). Изучение фразеологи и с использованием нашего проекта существенно меняет и представления о времени вхождения тех или иных фразеологических единиц в русский язык.

При просмотре корпуса примеров на отдельные лексические единицы по материалам «Библиотеки лексикографа» наглядно и с большой полнотой вскрываются цитаты из каких-либо текстов в других сочинениях, аллюзии и реминисценции. Такая задача является второстепенной для лингвистов, однако возможные наблюдения в этой области будут иметь существенное значение для изучения топики в перспективе истории русского литературного языка XVIII – начала XXI веков.

Изучение ономастики с его помощью даёт в распоряжение пользователей ономастический тезаурус, поскольку аналогичные источники картотечного типа отсутствуют вообще. Применение «Библиотеки лексикографа» для работы с именами собственными позволяет успешно решать следующие задачи, являющиеся весьма сложными в плане выбора материала:

1) поиск редких антропонимов и изучение их истории;
2) поиск вариантов гипокористических форм личных имён;
3) поиск форм отчеств в формах мужского и женского рода и в нетривиальных формах типа кузьмичи;
4) поиск фразеологизмов с антропонимами, этнонимами, топонимами;
5) поиск этнонимов в текстах разных жанров;
6) изучение различных форм этнонимов, в том числе устаревших, просторечных, региональных;
7) изучение различных дериватов от этнонимов (формы женского рода, названия детей);
8) поиск топонимов в текстах разных жанров;
9) классификация микротекстов с топонимами.

В отличие от словарей и подавляющего большинства картотек, ресурс «Библиотека лексикографа», дающий в распоряжение исследователя практически полную выборку примеров, позволяет изучать аксиологические характеристики лексических единиц и в том числе имён собственных – топонимов, этнонимов, антропонимов. С его помощью можно изучать интертекстуальные связи, выстраиваемые на основе имён собственных, выявлять скрытые цитаты, реминисценции, единицы топики на основе имён собственных разных разрядов.

Работа над любым массивом текстового материала неизбежно соприкасается с вопросами графической репрезентации текста. Понятно, что решение этих вопросов во многом зависит от задач, на решение которых направлен корпус или массив текстов: так, ориентация корпуса на отражение современного состояния языка жёстко требует выдерживания и единообразия графико-орфографических норм современного русского языка. Положение дел ещё сложнее, если корпус, как «Библиотека лексикографа», разрабатываемая в Словарном отделе ИЛИ РАН должна охватывать тексты XVIII – начала XXI веков, при этом для разработчиков оказывается желательным, чтобы в ней учитывалась аутентичная орфография текстов, позволяющая наблюдать процессы изменения в русской орфографии за последние три столетия не по словарям и нормативным сводам правил, а по текстам. Добавим к этому, что для ряда периодов истории русского литературного языка, охватывающих целые десятилетия, нормативные орфографические словари вообще отсутствуют, и как изучать характер норм орфографии, например, в интервале 1918–1955 годов – не вполне понятно. Отдельный вопрос – как получить электронные версии текстов, свободные от приведения орфографии к современным стандартам и пристрастиям внутри этих стандартов, но он здесь выходит за рамки обсуждения.

Проблема «буквы Ё» становится камнем преткновения как для лингвистических исследований с помощью электронных версий текстов, так и для разработок шрифтов и программ для работы с текстами. Венцом всего здесь является «обезъяченный и объёшенный» словарь В. И. Даля в современной версии. Не только в нём, но и, например, в электронных версиях словаря Д. Н. Ушакова и С. И. Ожегова не только расставлена буква «ё», отсутствующая в исходных печатных версиях этих словарей, но и введён новый порядок расположения слов – слова с «Ё» стоят после слов с «Е», а не вперемежку с ними, как традиционно принято в русской лексикографии. Активное внедрение графического знака «Ё» (буквой в строгом смысле слова он не является из-за особенностей функционирования и самого характера действующих норм, провозглашающих факультативность) приводит к следующему: в громадном количестве «объёшенных» текстов знак «Ё» выбивается из кодировки и отражается в виде непонятных символов, которые не идентифицируются при поиске и препятствуют выявлению слов с расставленным «Ё» в тексте. Поисковые инструменты большинства программ «разводят» даже читаемые знаки «Е» и «Ё», в результате чего, например, слово тетя приходится искать в двух вариантах – тетя и тётя, при этом ретроспективная расстановка буквы «Ё» в текстах XIX – первой трети XX вв. затемняет аутентичный графический облик этих текстов. По счастью, некоторые поисковые программы не делают разницы между «Е» и «Ё» и выдают по-разному оформленные слова в едином перечне ответов на запрос, но это имеет место лишь тогда, когда знак «Ё» идентифицируется данной программой – а это отмечается не всегда.

Для решения данных проблем при пополнении корпуса и его редактировании нам приходится внимательно просматривать новый текст и заменять нечитаемые символы, проявляющиеся вместо «Ё», или самое «Ё» знаком «Е». Это выполнимо, однако составляет трудности на большом объёме текстов.

Слитные, дефисные и раздельные написания, в особенности первые две названные группы, довольно непротиворечиво выявляются в корпусе при поиске слов и словосочетаний, причём полученные наблюдения оказываются информативными: так, нам удалось выявить более десятка написаний слова подшофе в раздельном, дефисном и слитном оформлении, которые вполне корреспондируют нормам отдельных периодов истории русской орфографии и наглядно иллюстрируют их вариантность.

Другая проблема, приобретающая особую важность в последнее время, – это представление русских текстов в аутентичной графике и орфографии или в графике и орфографии, действовавшей до 1917 года включительно, компенсирующее «обезъяченные» версии XX века или воспроизводящее тексты, не переиздававшиеся после 1918 года. Стремительно растёт число таких текстов в Библиотеке Мошкова, являющейся ценным источником для текстового материала. Такие тексты очень ценны для исследования русской лексики. Однако их появление сопряжено с целым клубком сложностей.

Знаки дореволюционной гражданской кириллицы – «ять», «фита», «ижица», даже если и отражаются аутентично в скопированных и интегрированных в корпус текстах, но не распознаются поисковыми программами, имеющими собственные заданные заранее шрифтовые настройки. Буква «і» распознаётся программами, но возникают проблемы с идентификацией слов и форм со стороны морфологии: ни один известный морфологический анализатор не работает с русской морфологией в старой орфографии (это касается и форм с конечным «ъ»). Каждый очередной текст в старой русской орфографии побуждает к размышлениям: заменить отменённые в 1918 году буквы и привести текст к современной орфографии или дождаться появления более совершенной программы, которая будет нивелировать различия между дореформенным и послереформенным написаниями слов. Впрочем, тут встречаются подлинные филологические «шедевры» – например, тексты произведений В. И. Даля, где сохраняются «ять», «ъ», «і», и тому подобное, но при этом оказывается расставленным «Ё», отсутствующее в оригинальных текстах[1].

Изучение орфографических вариантов слов, относящихся к периодам 1920–1950 годов по корпусу текстов и по «Библиотеке лексикографа», безусловно, сопряжено с рядом сложностей. Во-первых, далеко не все тексты этого периода существуют в электронном виде в авторской орфографии: априори чаще всего в них представлена орфография последнего издания, хотя есть возможность вводить в корпус отдельные тексты, преобразованные из форматов PDF и Djvu. Во-вторых, для того, чтобы обнаруживать те или иные написания в корпусе, надо иметь их список, который пока в обороте отсутствует. Тем не менее корпус и здесь может составить альтернативу классической словарной картотеке, поскольку мы не располагаем данными, в каком объёме в указанный период пополнялась Большая словарная картотека и сохранялись ли в картотеке авторские написания слов – мы знаем точно, что при составлении словарей написания слов, данные в текстовых источниках и с некоторой вероятностью имеющиеся в картотеке, приводились к нормам действующей на момент издания словаря орфографии.

---

[1] Только в наши дни понятно, что разработка русских шрифтов вполне могла могла бы базироваться не на современном русском алфавите, а на дореволюционном алфавите, в который, как известно, реформа 1918 года не добавила ничего, кроме апострофа вместо «Ъ» внутри слова. То есть во всех компьютерах мог бы использоваться комплект знаков, соответствующий современной русской графике, но те же шрифты позволяли бы и читать, и редактировать тексты в дореволюционной орфографии и большинство древнерусских текстов в упрощённой графике, а равно и преобразовывать тексты из одной формы русской орфографии в другую без потери шрифтовой идентичности. В настоящее время составление корпуса древнерусских текстов составляет почти бесперспективную задачу именно в силу того, что представляется нереальным оперативно привести имеющиеся в достаточном количестве тексты XI–XVII веков к единому графическому облику.

Так или иначе исследование графико-орфографических вариантов слов при помощи корпуса текстов или «Библиотеки лексикографа» намного – на несколько порядков – увеличивает объём доступного материала, хотя и оставляет желать много лучшего в отношении исходных данных. Впрочем, исторические словари русской орфографии как жанр в отечественной лексикографической традиции пока отсутствуют.

Параллельное, одновременное использование ресурсов «Библиотеки лексикографа» и классической по форме картотеки – Большой словарной картотеки – вызвало к жизни ряд вопросов, относящихся к сравнению стандартной бумажной картотеки и электронной картотеки в лексикологических исследованиях и лексикографической практике. В первую очередь эти вопросы касаются структуры самих информационных ресурсов, о чём отчасти сказано выше, отчасти в наших предшествующих публикациях. Следующая группа вопросов – это наличие и характер «внешней» информации в ресурсах, относящейся к конкретным источникам либо самим текстам, либо минимальным структурным единицам ресурсов. Если вопросы, связанные со справочным аппаратом и информационным обеспечением вновь создаваемых картотек могут решаться параллельно с работой над самими ресурсами, то положение с информационным обеспечением оказывается особенно сложным и сопряжённым либо с чрезвычайно трудными, либо с попросту нерешаемыми задачами.

Основная единица хранения информации в картотеке – карточка с текстовым или словарным примером. Основная единица хранения в «Библиотеке лексикографа» и любом корпусе текстов – это текст, конкретно текстовый файл. По объёму единиц, представляющих объект внимания составителя и пользователя, корпус текстов оказывается на два порядка компактнее картотеки, и, в то же время, по крайней мере, на три порядка превосходит картотеку по числу присутствующем в нём словоформ (в Большой Словарной картотеке 7 млн. карточек со словами (точнее, словоформами), в «Библиотеке лексикографа» – 3,1 млрд. словоформ, на каждую из которых в идеале при традиционной форме хранения и репрезентации лексикографических ресурсов должна была быть заведена отдельная бумажная карточка. Единицей пополнения корпуса является текст или документ как собрание текстов (напр., номер газеты или журнала), при этом каждый новый текст в корпусе обеспечивает доступ ко всем без исключения словам данного текста. В то же время даже для новых электронных картотек провозглашается принцип выборочности материала, не говоря о том, что в классических картотеках с многолетней историей, последствия работы нескольких поколений выборщиков в аспекте умышленных отбраковок (касающихся источников целиком или отдельных цитат из источников, неконформных в политическом или лингвокультурном отношении) или непредумышленных пропусков ценного материала по существу неустранимы.

Любая картотека строится по умолчанию на принципе «один источник – один документ на данный фрагмент текста», то есть – одна карточка. В составе «Библиотеки лексикографа» присутствуют тексты одних и тех же авторов в разном составе: отдельные романы, повести, рассказы, под-

борки стихотворений с датами создания и отдельные тома полных собраний сочинений или отдельные тома многотомных собраний сочинений (возможно размещение разных по составу собраний сочинений одного и того же автора, напр., Л. Н.Толстого, В. Г. Короленко и т. д.). Такой подход к источникам примеров обеспечивает надёжность отражения текстов в корпусе и предоставляет возможность давать ссылки либо на отдельное произведение (с указанием даты его создания), либо на том собрания сочинений того или иного автора. При таком подходе к материалу корпус по умолчанию включает, по крайней мере, часть доступных вариантов текстов произведений и позволяет следить за вариантами интересующих исследователя фрагментов текста.

Границы цитаты на бумажной или в электронной карточке жёстко раз и навсегда определены выборщиком: случаи правки текстов цитат в картотеках нам неизвестны. Границы извлекаемой цитаты при использовании корпуса или «Библиотеки лексикографа» определяет сам пользователь, копирующий необходимый объём текста, например, при применении программы Архивариус3000.

Ручная выборка иллюстративных материалов неизбежно тенденциозна и пристрастна: это касается как отбора самих источников, так и выбора цитат из них – невозможно выбрать вручную все употребления того или иного слова из одного конкретного текста, и это признается нецелесообразным. Отобранные вручную цитаты в лучшем случае иллюстрируют семантику слова: ни образный, ни аксиологический компоненты содержательной стороны (интенсионала) слова не могут быть представлены в объективированном виде ни в картотеке, ни тем более в словаре – в лучшем случае представленная в них картина будет отражать чьи-то субъективные представления, и это нельзя недооценивать (см. [Герд, 2008, 2011]). Корпус текстов и «Библиотека лексикографа», освобождённые от выборки – они могут страдать только от недостатка материалов, который легко компенсируется с течением времени – в этом плане дают максимально объективную картину. То же касается расстановки стилистических помет: по материалам картотек или системы словарей, используемых в работе, стилистическая характеристика слов достаточно субъективна: жанровая и тематическая разметка «Библиотеки лексикографа» позволяет выработать статистические критерии для расстановки стилистических помет и определения терминологического статуса слов.

Расположение иллюстративных материалов, документирующих одно и то же слово, в картотеке по существу ничем не регламентировано: в каких-то случаях расстановка карточек за обозначающим слово разделителем отражает какой-то этап работы над последним по времени словарём и группировку цитат по значениям. В корпусе и в «Библиотеке лексикографа» материал по умолчанию выстраивается по алфавиту авторов. При минимальной модификации рабочей версии Библиотеки была изготовлена новая версия, в которой материал аранжирован по датам создания текстов.

Расстановка дат создания текстов или в случае невозможности установления даты написания текстов – дат жизни их авторов (даже у М. Горько-

го имеется множество рассказов, не имеющих точной даты написания или прижизненной публикации), по опыту работы с «Библиотекой лексикографа» не составляет особого труда: эти даты вносятся в содержание текстового файла или в имя файла. Расстановка даты создания произведений на карточках многомиллионной картотеки – задача нереальная, выставление дат произведений в электронной картотеке – задача достаточно сложная.

Выгодным достоинством картотеки является наличие или возможность создания словника к ней, то есть указателя всех включённых в неё лексических единиц в исходной словарной форме. Однако работа над словником картотеки требует больших затрат сил и впоследствии – дополнительных забот по его пополнению и обновлению. Теоретически создание словника для корпуса текстов или его фрагментов представляет собой несложную задачу при применении широко распространённых программ, преобразующих тексты в списки словоформ (некоторые из программ подсчитывают частоту встречаемости словоформ и могут выдавать частотные или алфавитно-частотные списки словоформ), реально такая задача не имеет большого смысла.

Важный компонент любого источника словарных материалов. будь то картотека или корпус – это список источников материала. Каталог источников бумажной картотеки или «сократитель» словаря – как правило, печатный список названий текстов, обрастающий многочисленными дополнениями при пополнении картотеки. Каталог источников электронной картотеки может легко пополняться, но в любом случае он не столь доступен, как материалы картотеки. Каталог текстов «Библиотеки лексикографа» обновляется при любом её пополнении и извлекается из ресурса при помощи программ-каталогизаторов за одну минуту и несколько кликов мышью. В действующей версии Библиотеки этот каталог содержит даты создания подавляющего большинства текстов, не имеют дат в основном тексты, извлеченные из Интернета на ранних этапах комплектования ресурса.

Основное предназначение ресурса «Библиотека лексикографа» – это лексикология и лексикография. Однако в ресурс включена научная и научно-популярная литература по всем областям знаний и практик, что само по себе позволяет пользоваться им как информационно-поисковой системой по истории, географии, этнографии. Одновременно с реализацией данного проекта автором ведётся также собирание электронных словарей и справочников по всем областям знаний. Добавление в «Библиотеку лексикографа» электронных версий основных энциклопедий и энциклопедических словарей позволит использовать ресурс как информационно-поисковую систему, ориентированную на гуманитарные науки.

Список литературы

1. Баранов, А. Н. Введение в прикладную лингвистику [Текст] / А. Н. Баранов. – 3-е изд., стереот. – М. : Эдиториал УРСС, 2007. – 360 с.
2. Белозёрова, Н. Н. Шекспир и компания, или использование электронных библиотек при лингвистическом исследовании: учеб. пособие [Текст] / Н. Н. Бе-

лозёрова, Л. Е. Чуфистова ; М-во образования и науки Рос. Федерации, Тюмен. гос. ун-т. – Тюмень : Издательство Тюмен. гос. университета, 2007. – 296 с.

3. Богданова, С. Ю. Возможности корпусной методологии в решении лингвистических задач [Текст] / С. Ю. Богданова // Вестник Иркутского государственного лингвистического университета. – 2012. – Вып. 2. – С. 47–53.
4. Бурыкин, А. А. О создании электронной библиотеки для исследований в области русской лексикологии и лексикографии «Библиотека лексикографа» [Текст] / А. А. Бурыкин // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. El'Manuscript-08: материалы Междунар. науч. конф., Казань, 26–30 авг. 2008 г. – Казань : Изд-во КГУ, 2008. – С. 52–55.
5. Бурыкин, А. А. Электронная библиотека для исследований в области русской лексикологии и лексикографии: опыт работы, перспективы пополнения, возможности использования [Текст] / А. А. Бурыкин // Информационные технологии и письменное наследие. El'manuscript-10: материалы Междунар. науч. конф., Уфа, 28–31 октября 2010 г. – Уфа, Ижевск, 2010. – С. 36–41.
6. Бурыкин, А. А. Проблемы и задачи справочного аппарата к корпусам и картотекам [Электронный ресурс] / А. А. Бурыкин // Труды междунар. конф. «Корпусная лингвистика-2013», Санкт-Петербург, 25–27 июня 2013 г. – СПб., 2013. – С. 208–216. – Режим доступа : http://corpora.phil.spbu.ru/Works2013/Бурыкин.pdf
7. Бурыкин, А. А. Электронный ономастический тезаурус: проект «Библиотека лексикографа» в приложении к ономастике [Текст] / А. А. Бурыкин // Ономастика Поволжья: материалы XIV Междунар. науч. конф., Тверь, 10–12 сентября 2014 г. – Тверь, 2014. – С. 25–30.
8. Герд, А. С. Национальный корпус русского языка – словарная картотека – академический словарь [Электронный ресурс] / А. С. Герд // Труды междунар. конф. «Корпусная лингвистика-2008». – СПб., 2008. – С.143–148. – Режим доступа : http://corpora.phil.spbu.ru/Works2008/Gerd_143_148.pdf
9. Герд, А. С. Академическая лексикография как система корпусов [Электронный ресурс] / А. С. Герд // Труды междунар. конф. «Корпусная лингвистика-2013». – СПб., 2013. – С. 247–249. – Режим доступа : http://corpora.phil.spbu.ru/Works2013/Герд.pdf
10. Герд, А. С. Корпус текстов и источниковедение [Электронный ресурс] / А. С. Герд // Труды междунар. конф. «Корпусная лингвистика-2011». – СПб., 2011. – С.120–124.– Режим доступа : http://corpora.phil.spbu.ru/Works2011/Герд_120.pdf
11. Герд, А. С. Национальный корпус русского языка в свете проблем современной филологии [Электронный ресурс] / А. С. Герд, В. П. Захаров // Труды международной конференции «Корпусная лингвистика-2004». – СПб., 2004. – С. 122–131. – Режим доступа : http://corpora.phil.spbu.ru/Works2004/Gerd_Zakharov_art.pdf
12. Глинкина, Л. А. Словарь забытых и трудных слов из произведений русской литературы XVIII–XIX веков [Текст] / Л. А. Глинкина. – Оренбург : Оренбургское книжное изд-во, 1998. – 277 с.
13. Грудева, Е. В. Корпусная лингвистика [Текст] / Е. В. Грудева. – 2-е изд., стер. – М. : ФЛИНТА, 2012. – 165 с.
14. Елистратов, В. С. Язык старой Москвы: лингвоэнциклопедический словарь [Текст] / В. С. Елистратов. – М. : Русские словари, 1997. – 795 с.

15. Захаров, В. П. Корпусная лингвистика: учеб.-метод. пособие [Текст] / В. П. Захаров. – СПб., 2005. – 48 с.
16. Захаров, В. П. Корпусная лингвистика : учеб. для студентов гуманит. вузов [Текст] / В. П. Захаров, С. Ю. Богданова. – Иркутск : ИГЛУ, 2011. – 161 с.
17. Захаров, В. П. Словарная картотека Института лингвистических исследований как объект автоматизации [Электронный ресурс] / В. П. Захаров // Компьютерная лингвистика и интеллектуальные технологии : труды междунар. конф. «Диалог–2007». – М., 2007. – Режим доступа : http://www.dialog-21.ru/digests/dialog2007/materials/html/31.htm
18. Зубов, А. В. Информационные технологии в лингвистике [Текст] : учеб. пособие / А. В. Зубов, И. И. Зубова. – М. : Академия, 2004. – 208 с.
19. Копотев, М. В. К построению частотной грамматики русского языка [Электронный ресурс] / М. В. Копотев // Труды международной конференции «Корпусная лингвистика-2008». – СПб., 2008. – С. 207–213. – Режим доступа : http://corpora.phil.spbu.ru/Works2008/Kopotev_207_213.pdf
20. Национальный корпус русского языка: 2003–2005. Сборник статей [Электронный ресурс]. – М. : Индрик, 2005. Режим доступа : http://ruscorpora.ru/
21. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы [Текст] / отв. ред. В. А. Плунгян. – СПб. : Нестор-История, 2009. – 502 с.
22. Перцов, Н. В. О роли корпусов в лингвистических исследованиях [Электронный ресурс] / Н. В. Перцов // Труды междунар. конф. «Корпусная лингвистика-2006». – СПб., 2006. – С. 318–331. – Режим доступа : http://corpora.phil.spbu.ru/Works2006/Pertsov_doklad_318_331.pdf
23. Плисецкая, А. Д. Национальный корпус русского языка как один из инструментов анализа фразеологических сочетаний [Электронный ресурс] / А. Д. Плисецкая // Труды междунар. конф. «Корпусная лингвистика-2013». – СПб., 2013. – С. 387–396. – Режим доступа : http://corpora.phil.spbu.ru/Works2013/Плисецкая.pdf
24. Редкие слова в произведениях авторов XIX века: слов.-справ. : ок. 3500 единиц [Текст] / отв. ред. Р. П. Рогожникова. – М. : Русские словари, 1997. – 572 с.
25. Рогожникова, Р. П. Практическая лексикография. 100 лет словарной картотеке [Текст] / Р. П. Рогожникова. – М. : 1989. – 126 с.
26. Рогожникова, Р. П. Школьный словарь устаревших слов русского языка: По произведениям русских писателей XVIII–XX вв. [Текст] / Р. П. Рогожникова, Т. С. Карская – М. : Просвещение: Учебная литература, 1996. – 608 с.
27. Рогожникова, Р. П. Сокровищница русского слова: история большой словарной картотеки Института лингвистических исследований РАН [Текст] / Р. П. Рогожникова [отв. ред. Н. Н. Казанский]. – СПб. : Наука, 2003. – 106 с.
28. Русская судьба крылатых слов [Текст] / отв. ред. В. Е. Багно. – СПб. : Наука, 2010. – 664 с.
29. Соловьев, В. Д. Частотность как объект корпусных исследований [Электронный ресурс] / В. Д. Соловьев // Труды междунар. конф. «Корпусная лингвистика-2011». – СПб., 2011. – С. 328–332. Режим доступа : http://corpora.phil.spbu.ru/Works2011/Соловьев_328.pdf
30. Сомов, В. П. Словарь редких и забытых слов [Текст] / В. П. Сомов. – М. : Владос, 1996. – 764 с.

31. Тетакаева, Л. М. Учебно-методический комплекс по дисциплине «Корпусная лингвистика» [Электронный ресурс] / Л. М. Тетакаева. – Махачкала, 2011. – 16 с. – Режим доступа : http://fia.dgu.ru/Content/files/УМК/Корпусная20%лингвистика.pdf

References

1. Baranov, A. N. Vvedenie v prikladnuju lingvistiku [Tekst] / A. N. Baranov. – 3-e izd., stereot. – M. : Jeditorial URSS, 2007. – 360 s.
2. Belozjorova, N. N. Shekspir i kompanija, ili ispol'zovanie jelektronnyh bibliotek pri lingvisticheskom issledovanii: ucheb. posobie [Tekst] / N. N. Belozjorova, L. E. Chufistova ; M-vo obrazovanija i nauki Ros. Federacii, Tjumen. gos. un-t. – Tjumen' : Izdatel'stvo Tjumen. gos. universiteta, 2007. – 296 s.
3. Bogdanova, S. Ju. Vozmozhnosti korpusnoj metodologii v reshenii lingvisticheskih zadach [Tekst] / S. Ju. Bogdanova // Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo universiteta. – 2012. – Vyp. 2. – S. 47–53.
4. Burykin, A. A. O sozdanii jelektronnoj biblioteki dlja issledovanij v oblasti russkoj leksikologii i leksikografii «Biblioteka leksikografa» [Tekst] / A. A. Burykin // Sovremennye informacionnye tehnologii i pis'mennoe nasledie: ot drevnih tekstov k jelektronnym bibliotekam. El'Manuscript-08: materialy Mezhdunar. nauch. konf., Kazan', 26–30 avg. 2008 g. – Kazan' : Izd-vo KGU, 2008. – S. 52–55.
5. Burykin, A. A. Jelektronnaja biblioteka dlja issledovanij v oblasti russkoj leksikologii i leksikografii: opyt raboty, perspektivy popolnenija, vozmozhnosti ispol'zovanija [Tekst] / A. A. Burykin // Informacionnye tehnologii i pis'mennoe nasledie. El'manuscript-10: materialy Mezhdunar. nauch. konf., Ufa, 28–31 oktjabrja 2010 g. – Ufa, Izhevsk, 2010. – S. 36–41.
6. Burykin, A. A. Problemy i zadachi spravochnogo apparata k korpusam i kartotekam [Jelektronnyj resurs] / A. A. Burykin // Trudy mezhdunar. konf. «Korpusnaja lingvistika-2013», Sankt-Peterburg, 25–27 ijunja 2013 g. – SPb., 2013. – S. 208–216. – Rezhim dostupa : http://corpora.phil.spbu.ru/Works2013/Burykin.pdf
7. Burykin, A. A. Jelektronnyj onomasticheskij tezaurus: proekt «Biblioteka leksikografa» v prilozhenii k onomastike [Tekst] / A. A. Burykin // Onomastika Povolzh'ja: materialy XIV Mezhdunar. nauch. konf., Tver', 10–12 sentjabrja 2014 g. – Tver', 2014. – S. 25–30.
8. Gerd, A. S. Nacional'nyj korpus russkogo jazyka – slovarnaja kartoteka – akademicheskij slovar' [Jelektronnyj resurs] / A. S. Gerd // Trudy mezhdunar. konf. «Korpusnaja lingvistika-2008». – SPb., 2008. – S. 143–148. – Rezhim dostupa : http://corpora.phil.spbu.ru/Works2008/Gerd_143_148.pdf
9. Gerd, A. S. Akademicheskaja leksikografija kak sistema korpusov [Jelektronnyj resurs] / A. S. Gerd // Trudy mezhdunar. konf. «Korpusnaja lingvistika-2013». – SPb., 2013. – S. 247–249. – Rezhim dostupa : http://corpora.phil.spbu.ru/Works2013/Gerd.pdf
10. Gerd, A. S. Korpus tekstov i istochnikovedenie [Jelektronnyj resurs] / A. S. Gerd // Trudy mezhdunar. konf. «Korpusnaja lingvistika-2011». – SPb., 2011. – S. 120–124. – Rezhim dostupa : http://corpora.phil.spbu.ru/Works2011/Gerd_120.pdf
11. Gerd, A. S. Nacional'nyj korpus russkogo jazyka v svete problem sovremennoj filologii [Jelektronnyj resurs] / A. S. Gerd, V. P. Zaharov // Trudy mezhdunarodnoj

konferencii «Korpusnaja lingvistika-2004». – SPb., 2004. – S. 122–131. – Rezhim dostupa : http://corpora.phil.spbu.ru/Works2004/Gerd_Zakharov_art.pdf
12. Glinkina, L. A. Slovar' zabytyh i trudnyh slov iz proizvedenij russkoj literatury XVIII–XIX vekov [Tekst] / L. A. Glinkina. – Orenburg : Orenburgskoe knizhnoe izd-vo, 1998. – 277 s.
13. Grudeva, E. V. Korpusnaja lingvistika [Tekst] / E. V. Grudeva. – 2-e izd., ster. – M. : FLINTA, 2012. – 165 s.
14. Elistratov, V. S. Jazyk staroj Moskvy: lingvojenciklopedicheskij slovar' [Tekst] / V. S. Elistratov. – M. : Russkie slovari, 1997. – 795 s.
15. Zaharov, V. P. Korpusnaja lingvistika: ucheb.-metod. posobie [Tekst] / V. P. Zaharov. – SPb., 2005. – 48 s.
16. Zaharov, V. P. Korpusnaja lingvistika : ucheb. dlja studentov gumanit. vuzov [Tekst] / V. P. Zaharov, S. Ju. Bogdanova. – Irkutsk : IGLU, 2011. – 161 s.
17. Zaharov, V. P. Slovarnaja kartoteka Instituta lingvisticheskih issledovanij kak ob#ekt avtomatizacii [Jelektronnyj resurs] / V. P. Zaharov // Komp'juternaja lingvistika i intellektual'nye tehnologii : trudy mezhdunar. konf. «Dialog–2007». – M., 2007. – Rezhim dostupa : http://www.dialog-21.ru/digests/dialog2007/materials/html/31.htm
18. Zubov, A. V. Informacionnye tehnologii v lingvistike [Tekst] : ucheb. posobie / A. V. Zubov, I. I. Zubova. – M. : Akademija, 2004. – 208 s.
19. Kopotev, M. V. K postroeniju chastotnoj grammatiki russkogo jazyka [Jelektronnyj resurs] / M. V. Kopotev // Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika-2008». – SPb., 2008. – S. 207–213. – Rezhim dostupa : http://corpora.phil.spbu.ru/Works2008/Kopotev_207_213.pdf
20. Nacional'nyj korpus russkogo jazyka: 2003–2005. Sbornik statej [Jelektronnyj resurs]. – M. : Indrik, 2005. Rezhim dostupa : http://ruscorpora.ru/
21. Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy [Tekst] / otv. red. V. A. Plungjan. – SPb. : Nestor-Istorija, 2009. – 502 s.
22. Percov, N. V. O roli korpusov v lingvisticheskih issledovanijah [Jelektronnyj resurs] / N. V. Percov // Trudy mezhdunar. konf. «Korpusnaja lingvistika-2006». – SPb., 2006. – S. 318–331. – Rezhim dostupa : http://corpora.phil.spbu.ru/Works2006/Pertsov_doklad_318_331.pdf
23. Pliseckaja, A. D. Nacional'nyj korpus russkogo jazyka kak odin iz instrumentov analiza frazeologicheskih sochetanij [Jelektronnyj resurs] / A. D. Pliseckaja // Trudy mezhdunar. konf. «Korpusnaja lingvistika-2013». – SPb., 2013. – S. 387–396. – Rezhim dostupa : http://corpora.phil.spbu.ru/ Works2013/Pliseckaja.pdf
24. Redkie slova v proizvedenijah avtorov XIX veka: slov.-sprav. : ok. 3500 edinic [Tekst] / otv. red. R. P. Rogozhnikova. – M. : Russkie slovari, 1997. – 572 s.
25. Rogozhnikova, R. P. Prakticheskaja leksikografija. 100 let slovarnoj kartoteke [Tekst] / R. P. Rogozhnikova. – M. : 1989. – 126 s.
26. Rogozhnikova, R. P. Shkol'nyj slovar' ustarevshih slov russkogo jazyka: Po proizvedenijam russkih pisatelej XVIII–XX vv. [Tekst] / R. P. Rogozhnikova, T. S. Karskaja – M. : Prosveshhenie: Uchebnaja literatura, 1996. – 608 s.
27. Rogozhnikova, R. P. Sokrovishhnica russkogo slova: istorija bol'shoj slovarnoj kartoteki Instituta lingvisticheskih issledovanij RAN [Tekst] / R. P. Rogozhnikova [otv. red. N. N. Kazanskij]. – SPb. : Nauka, 2003. – 106 s.

28. Russkaja sud'ba krylatyh slov [Tekst] / otv. red. V. E. Bagno. – SPb. : Nauka, 2010. – 664 s.
29. Solov'ev, V. D. Chastotnost' kak ob#ekt korpusnyh issledovanij [Jelektronnyj resurs] / V. D. Solov'ev // Trudy mezhdunar. konf. «Korpusnaja lingvistika-2011». – SPb., 2011. – S. 328–332. Rezhim dostupa : http://corpora.phil.spbu.ru/Works2011/Solov'ev_328.pdf
30. Somov, V. P. Slovar' redkih i zabytyh slov [Tekst] / V. P. Somov. – M. : Vlados, 1996. – 764 s.
31. Tetakaeva, L. M. Uchebno-metodicheskij kompleks po discipline «Korpusnaja lingvistika» [Jelektronnyj resurs] / L. M. Tetakaeva. – Mahachkala, 2011. – 16 s. – Rezhim dostupa : http://fia.dgu.ru/Content/files/UMK/Korpusnaja20%lingvistika.pdf