О создании электронной библиотеки для исследований в области русской лексикологии и лексикографии «Библиотека лексикографа»

А. А. Бурыкин

Институт лингвистических исследований РАН, Санкт-Петербург, Россия

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 в данном случае оказывается, как мы полагаем, доступность материала, т.е. возможность его использования в каких угодно условиях, начиная от Интернета или компьютерной сети крупного научного центра до персонального компьютера студента-первокурсника.

В настоящее время наиболее известным и широко доступным источником русских текстов являются электронные библиотеки, доступные в Интернете (Библиотека Мошкова, Альдебаран и некоторые другие), те же библиотеки, выпущенные на компакт-дисках (Библиотека «Всемирная литература», Библиотека Мошкова с частью ресурсов), сайты, посвященные творчеству отдельных писателей, электронные собрания сочинений отдельных писателей XIX-XX веков, тематические собрания текстов (так, выпущен диск с образцами русской драматургии). Однако эти продукты, несмотря на всю полезность, имеют ряд недостатков технического порядка. Тексты, извлекаемые из электронных изданий, с большим трудом поддаются объединению, так как для этого требуется их дополнительная обработка. Электронные библиотеки как в ресурсах Интернета, так и на компакт-дисках, сильно разнятся по своему составу, переводная художественная литература в них преобладает над отечественной (а в последней – современные боевики и детективы доминируют над классикой), а единая для диска поисковая система (такая имеется в Библиотеке «Всемирная литература») при применении ее к лексическому материалу хотя и действует, но оказывается бесполезной.

Необходимость дальнейшей оптимизации работы в области русской лексикологии и лексикографии как синхронной, 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 адресована пользователям-филологам.

Данный проект, получивший название «Библиотека лексикографа», реализуется автором в словарном отделе ИЛИ РАН с начала 2008 г. Источником материала являются все доступные в Интернете электронные библиотеки, а также собрания текстов на компакт-дисках. Материал для библиотеки отбирается примерно по тем же принципам, по которым комплектовалась и продолжает комплектоваться Библиотека Словарного отдела ИЛИ РАН, являющаяся хранилищем источников для Большой словарной картотеки ИЛИ РАН. Хронология охвата материала – от начала XVIII века до начала XXI века (ведется собирание древнерусских текстов XI-XVII вв. в электронном виде, но этот материал имеет ограниченное применение в связи с различиями в техническом оформлении текста и передаче графики). 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 т п.;
- научно-популярная литература и учебные пособия по всем областям знаний.

Автором ведется также сбор электронных словарей и справочников по всем областям знаний, а также сбор лингвистической литературы в электронной форме, однако пока эти тексты непосредственно в

«Библиотеке лексикографа» не размещаются и составляют отдельные коллекции.

Какими достоинствами обладает проект «Библиотека лексикографа» по сравнению с иными ресурсами, содержащими русские тексты и лексический материал русского языка?

Прежде всего это независимость от подключения к Интернету – отнюдь не везде и далеко не все исследователи имеют возможность часами эксплуатировать Интернет в поисках материалов 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 электронных библиотек имеет относительно успешное решение. В сфере художественной литературы тексты писателей русского зарубежья (М. Осоргин, Г. Газданов, Н. Берберова и др.), опальных отечественных писателей (Е. Замятин, Б. Пильняк, А. Веселый и др.), а также политических и военных деятелей (Л. Троцкий, П. Врангель, А. Деникин и др.) восполняют лакуну в отечественных лексикографических ресурсах, поскольку сочинения этих авторов никогда не привлекались для пополнения картотек. Имеется возможность разместить в максимально полном виде произведения таких авторов, как М. Булгаков, М. Зощенко, М. Цветаева, А. Ахматова, О. Мандельштам, Б. Пастернак, Б. Окуджава и других, чьи тексты в лексикографической работе почти не использовались или по крайней мере их воспроизведение в словарях не приветствовалось. Автором ставится задача не только подготовить качественно новый продукт, предназначенный для лексикографических работ и лексикологических исследований, но и за счет доступных ресурсов воссоздать в электронном виде корпус источников имеющихся словарей и, соответственно, материал Большой словарной картотеки ИЛИ РАН.

Переоценка ценностей в области русской литературы XX века приводит к тому, что произведения некоторых авторов, составлявшие

основу для первого издания 17-томного словаря современного русского языка (романы С. Бабаевского «Кавалер Золотой звезды»; М. Бубеннова «Белая береза» и т.п.) разыскиваются с большим трудом, но все же занимают свое место в Библиотеке. В ней присутствуют – рядом с книгами современных политиков – и «Краткий курс истории ВКП (б)», и работы В. И. Ленина и И. В. Сталина, и сочинения Л. И. Брежнева. Кстати, выбор текстов из электронных библиотек требует основательного знания «советской» литературы 1930-х – 1980-х 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 некоторые другие, то есть те, чье 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) 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 (например, «Угол падения» и др.) в электронном виде отсутствуют. Из сочинений В. Ажаева имеется роман «Вагон», более ранние и известные произведения этого автора («Далеко от Москвы» и др.) в Интернете не обнаруживаются. Не без труда отыскиваются ранние произведения В. Аксенова («Коллеги»), Д. Гранина («Искатели», «Иду на грозу»), редкостью оказываются произведения В. Пановой, а сочинения Г. Николаевой и В. Кетлинской попросту отсутствуют в электронных библиотеках.

Обработка электронных материалов, извлекаемых из библиотек Интернета, включает приведение их к единому формату TXT. Для пользования библиотекой применяется редактор Bred, позволяющий работать с txt-файлами любого объема. Поисковая система Integra (имеется в ресурсах Интернета и устанавливается на любой компьютер)

позволяет просматривать до 1000 цитат с необходимыми запрашиваемыми словами с учетом всех словоформ, она же позволяет копировать из Библиотеки корпус документов, где встречаются запрошенные слова. Возможно, более удобной поисковой программой окажется система Archivarius 3000. Из технических требований к компьютерам для работы с библиотеками наиболее значимы не столько объем жесткого диска, сколько быстродействие процессора и объем оперативной памяти.

В настоящее время «Библиотека лексикографа» включает в себя более 8000 текстов (объем около 1,7 Гб) и постоянно пополняется. Кроме обобщающего корпуса текстов в дальнейшем предполагается выделить в нем 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. 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 лексическую стилистику и язык писателей, наконец, данное собрание может быть «настроено» для составления толкового словаря русского языка любого типа, объема и хронологического диапазона.

On the creation of the Lexicographer’s Library, a digital library for research in the area of Russian lexicology and lexicography

Alexey A. Burykin
Institute for Linguistic Research of the Russian Academy of Sciences, St. Petersburg, Russia

This paper is a preliminary report on a project to develop the Lexicographer's Library, an electronic resource for studying the lexicology of Russian and lexicographical practices. The main part of the project (already been developed in the Lexicographic Department of the Institute for Linguistic Research of the Russian Academy of Sciences, St. Petersburg) is the collection of Russian texts (fiction, memoirs, scientific literature, political literature, laws, etc.) in electronic form to allow searching for words and idioms. The collection already contains more than 8,000 files representing Russian books from the 18th to the beginning of 21st centuries and continues to grow.