А. А. Бурыкин
A. A. Burykin

ПРОБЛЕМЫ И ЗАДАЧИ СПРАВОЧНОГО АППАРАТА К КОРПУСАМ И КАРТОТЕКАМ

PROBLEMS AND TASKS FOR THE REFERENCE TOOLS TO THE CORPUS MASSES AND CARD STORES

Аннотация. Темой работы являются вопросы построения справочного аппарата к корпусам текстов и различным картотекам, как традиционным бумажным картотекам, так и электронным собраниям примеров. Автор отмечает, что справочный аппарат к любому электронному собранию текстов более удобен с технологической точки зрения, более оперативен для отражения изменений в данных к текстам (биографии авторов, хронология и т.д.) и может иметь ценность самостоятельного биобиблиографического справочного ресурса.

Abstract. The theme of the Work are the questions of construction of the auxiliary resources to text corpuses and various card stores, both traditional paper card stores, and electronic collections of examples. The author notices that the auxiliary resources to any electronic collection of texts is more convenient from the technological point of view, it is more operative for introducing of changes, additions and corrections in the data to texts (to the biographies of authors, chronology etc.) and it even can have value of an independent biobibliographic auxiliary resource.

1. Общие положения

Проблемы справочного аппарата в приложении к большим массивам текстов в виде корпуса или к значительным объемам лексических материалов в форме электронных картотек являются не менее актуальными, чем работа над вспомогательным аппаратом бумажных картотек. Особое значение имеет справочный аппарат электронных ресурсов, ориентированных на лексикографическую работу, поскольку этот аппарат частично воспроизводится в словарях и по своему составу, отражающему

структуру ресурсов, влияет на качество проработки лексического материала в словаре в сравнении со словарями и исследованиями, разрабатываемыми на основе иных ресурсов и источников.

Настоящая работа представляет опыт работы над проектом «Библиотека лексикографа» собранием русских текстов для лексикологических исследований и лексикографической практики, который создан и используется в Словарном отделе Института лингвистических исследований РАН с 2008 г. В настоящее время «Библиотека лексикографа» включает более 42 тыс. текстов разных жанров и различной тематики и некоторое количество материалов из периодической печати. Суммарный объем ресурса – около 1,7 млрд словоформ. В группе Большого академического словаря ресурс используется для поиска материалов, для которых отсутствует или является недостаточной документация в Большой словарной картотеке ИЛИ РАН, в Группе новых слов – для поиска лексических единиц, получающих фиксацию как «новые слова» по данным современной периодики, в источниках более раннего времени. Подобная задача сделала актуальным такой параметр текстов, как **дата создания** текста, с обеспечением мгновенного доступа к этой информации. Эти данные введены в Библиотеке Лексикографа в версии 2013 года в имена файлов, просматриваемые при использовании проекта с программой Архивариус 3000 и сохранены в тех файлах, в которых они были проставлены при тексте (таковы все тексты, взятые из Библиотеки Мошкова).

Стремительно растущий объем библиотеки поставил задачу **жанровой и предметно-тематической классификации** текстов, для чего была разработана система индексов, вводимых в имена файлов. Эти индексы позволят автоматически выбирать из объема Библиотеки тексты, укладывающиеся по времени создания в периоды, равные трети века, в хронологическом интервале Библиотеки – XVIII-начало XXI вв. в объемах, 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 объем 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.

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 чем 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 информационные обеспечением вновь создаваемых картотек могут решаться параллельно с работой над самими ресурсами, то положение с информационным обеспечением оказывается особенно сложным и сопряженным либо с чрезвычайно трудными, либо с попросту нерешаемыми задачами.

2. Проблема единиц описания

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 объему единиц составляющих 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 нем

словоформ (в Большой Словарной картотеке 7 млн карточек со словами, в Библиотеке лексикографа – 1,7 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 (например, 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.

«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 (например, более ранние или более поздние, тексты определенных 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 объеме (3–4 тыс. примеров), например, при появлении более совершенного «академического» издания сочинений какого-либо автора по существу невозможно.

Любая картотека строится по умолчанию на принципе «один источник – один документ на данный фрагмент текста», т.е. одна карточка. В составе «Библиотеки лексикографа» присутствуют

тексты одних и тех же авторов в разном составе: отдельные романы, повести, рассказы, подборки 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, например Л.Н.Толстого, В.Г.Короленко и т.д.). такой подход к источникам примеров обеспечивает надежность отражения текстов в корпусе и дает 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 жестко 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 объем текста, например, при применении программы Архивариус3000.

3. Выборки или тезаурус? Технологичность и полнота как преимущества корпуса

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. Корпус

текстов и Библиотека лексикографа, освобожденные 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 словарем и группировку цитат по значениям. В корпусе и в Библиотеке лексикографа материал по умолчанию выстраивается по алфавиту авторов. При минимальной модификации рабочей версии библиотеки материал может быть аранжирован по датам создания текстов.

Расстановка дат создания текстов или в случае невозможности установления даты написания текстов – дат жизни их авторов (даже у М.Горького 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 включенных в нее 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 распространенных 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.

4. Биобиблиографический аппарат к корпусу как отдельный компьютерный продукт

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 пополнении библиотеки и извлекается из ресурса при помощи программ-каталогизаторов за одну минуту и несколько кликов мышью. В действующей версии Библиотеки этот каталог содержит даты создания большинства текстов.

Список источников материала любого ресурса будет неполным без биобиблиографических сведений об авторах, произведения которых присутствуют в корпусе или являются источниками картотеки. Такой вспомогательный ресурс к корпусу или картотеке должен удовлетворять следующим условиям: 1) полностью или в максимально полной мере охватывать персоналии, соотносительные с источниками; 2) содержать биографические и библиографические сведения об авторах – даты жизни, область занятий, названия сочинений, представленные в корпусе; 3) по возможность включать сведения о всех сочинениях данного автора: перечень сочинений, не

представленных в корпусе, составляет раздел desiderata для дальнейшей работы над ресурсом; 4) быть технологичным в обновлении, открытым для пользователей и тех, кто занят работой по пополнению ресурса и обновлению его аппарата; 5) учитывать новейшие биобиблиографические данные об авторах. Список источников к большой Словарной картотеке вообще не дает этой информации, в то время как она не только является необходимой как статический объем знаний, но и нуждается в постоянном мониторинге на предмет обновления, пополнения, уточнения.

Наиболее ценным источником биобиблиографической информации об авторах текстов, присутствующих в Библиотеке лексикографа, является Википедия – одна из наиболее новых электронных энциклопедий, материалы которой обновляются весьма часто и именно страницы Википедии обычно наиболее полно соответствуют тому или иному временному срезу (с ней не может сравниваться, например, Краткая литературная энциклопедия, отражающая данные 40-летней давности). Копирование страниц Википедии и других электронных ресурсов позволяет аккумулировать и хранить сведения об авторах и времени изданий большой части текстов. Ценным источником сведений об авторах XX века является также биобиблиографический словарь Игоря Авросимова (электронная версия на сайте Proza.ru), различные версии Большой русской Биобиблиографической энциклопедии, интегрировавшей более трех десятков биобиблиографических словарей, некоторые более специальные биобиблиографические справочники. Собирание этих изданий в электронной форме позволит в перспективе при минимуме затрат сил составить собственный биобиблиографический словарь-справочник к Библиотеке лексикографа и вести мониторинг биобиблиографических сведений с целью его обновления и приведения в соответствие с текущим моментом.