СЛОВАРНАЯ КАРТОТЕКА ИНСТИТУТА ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ РАН КАК ОБЪЕКТ АВТОМАТИЗАЦИИ[1]

DICTIONARY CARD FILES AS AN OBJECT FOR AUTOMATION

Захаров В.П. *(vz1311@yandex.ru)*
Институт лингвистических исследований РАН
Санкт-Петербургский государственный университет

В статье обсуждаются вопросы компьютеризации словарной картотеки ИЛИ РАН, насчитывающей около 8 млн. карточек. Обсуждается место и роль картотек и корпусов в работе лексикографа, а также создание специальных корпусов, ориентированных на создание словарей. Выдвигается идея формирования и ведения открытой картотеки в режиме online.

Проблемы автоматизации в лексикографии многоаспектны, и наиболее целесообразно обсуждать их применительно к конкретным типам словарей и видам лексикографических работ. Однако общепризнано, что в основе любого словаря лежит словарная картотека, откуда черпается необходимый лексический материал. При создании словарей национального языка только репрезентативная фундаментальная картотека, насчитывающая миллионы карточек, может считаться надежной базой. Источниками картотеки служат специально подобранные и обработанные тексты, отражающие язык во всем его многообразии.

Большая словарная картотека (БСК)[2] Института лингвистических исследований РАН (ИЛИ РАН), в которой собрано и систематизировано огромное количество словарных карточек с цитатами, позволяющих вести разнообразные словарные и филологические работы, насчитывает порядка 8 млн. карточек. Материалы картотеки были использованы при подготовке множества словарей и грамматик русского языка, включая такие фундаментальные труды, как «Словарь современного русского языка» в 17 тт., академическая «Грамматика русского языка» 1952-1954 гг., «Орфографический словарь русского языка» РАН и мн. др. [Рогожникова 1989, Рогожникова 2003, Приемышева 2003]. Картотекой пользуются ученые из различных городов нашей страны и из-за рубежа для исследований по различным проблемам русского языкознания. Сегодня с ее помощью создаются Большой академический словарь русского языка в 25 тт. [Большой академический словарь], новый фразеологический словарь русского языка и др.

Картотека начала создаваться еще в XIX веке под руководством академиков Я. К. Грота и А. А. Шахматова. В настоящее время Большая словарная картотека состоит из двух частей: старой картотеки (с 1886 по 1968 г.), насчитывающей около 5,5 млн. карточек, и новой картотеки (с 1968 по 1994 г.), насчитывающей около 2,5 млн. карточек. Наряду с БСК в ИЛИ РАН хранятся и другие картотеки, значительные по объему материалов (картотека Словаря русского языка XVIII века — около 3 млн. карточек, картотека Словаря русских народных говоров – более 2 млн. карточек, картотека Словаря новых слов и др.).

В 1986 и 2001 гг. было проведено две конференции, специально посвященных БСК. Среди всех выступлений особо стоит отметить высказывания, в которых подчеркивалось, что картотека должна охраняться государством как общенародное достояние, как памятник культуры.

Однако кроме "музейно-мемориальной" функции БСК и по сей день сохраняет и научное значение. Достаточно сказать, что лексический материал, собранный в картотеке превосходит все изданные словари русского языка. Если в печатных словарях (даже в больших) к отдельным словам приводится лишь несколько цитат – примеров употребления данного слова, то в картотеке таких цитат может быть несколько сотен. Но дело не только в богатстве иллюстративного материала. Как показали наши изыскания, все основные словари русского языка (см. сводный словник, составленный Р. П. Рогожниковой [Сводный словарь 1991]) покрывают не более 50% слов, зафиксированных в БСК.

Также на выше упомянутых конференциях говорилось о необходимости создания информационной базы современной академической лексикографии. В 1980-е гг. в рамках проекта создания Машинного фонда русского языка были начаты работы по автоматизации БСК, однако они были прерваны в самом начале. В настоящее время

[1] Программа ОИФН РАН на 2006–2008 гг. «Русский язык, литература и фольклор в информационном обществе: формирование электрон-ных научных фондов», проект «Развитие Большой словарной картотеки (БСК) Института лингвистических исследований РАН».

[2] Также существует название Большая картотека словарного отдела ИЛИ РАН (БКСО).

назрела настоятельная необходимость в компьютерной базе данных БСК и в ее пополнении на основе современных информационных технологий [Захаров 2005 б].

Прежде всего, база данных БСК призвана облегчить работу пользователей с картотекой и предоставить им новые возможности. Наличие базы картотеки должно также способствовать наведению в ней порядка. Большая картотека – это огромное количество карточек, хранящихся в каталожных ящиках в деревянных шкафах. За последнее время БСК трижды переезжала. После переездов и многолетних работ с картотекой выявляются нарушения в алфавитном порядке следования карточек, пропажа заставок слов и т.п. Кроме того, создание интегрированной базы данных картотеки позволит упорядочить форматы картотечных данных, создававшихся в разное время и по различным стандартам.

Автоматизации картотечных работ заключается в компьютерной поддержке технологических процессов пополнения и использования картотеки.

Это прежде всего:

• создание, пополнение картотеки;

• нахождение нужных карточек;

• работа с карточками в процессе написания словарных статей.

Последняя функция не имеет прямой связи с картотекой, а задача пополнения БСК делится, в свою очередь, на следующие подтипы:

• оптимизация и пополнение БСК с точки зрения репрезентативности;

• создание новых картотек (или пополнение БСК) применительно к задачам подготовки новых словарей.

Информационные массивы БСК представляют собой данные, порождаемые и используемые на различных этапах обеспечения процесса поддержки лексикографических работ. Прежде всего, это расставленные по алфавиту карточки слов-разделителей, за которыми следуют карточки с цитатами. Массив слов-разделителей может быть назван словником картотеки. Внутри словника существуют отсылки к другим словам и пометы (сведения об ударении, семантике и др.), которые тоже должны быть отражены в базе.

Второй массив, это собственно карточки с цитатами (см. рис. 1) и сопутствующей информацией, которую условно можно назвать “библиографической” и “технологической”. Первая представляет собой отсылки к источникам, из которых взяты цитаты, вторая содержит различную служебную информацию.

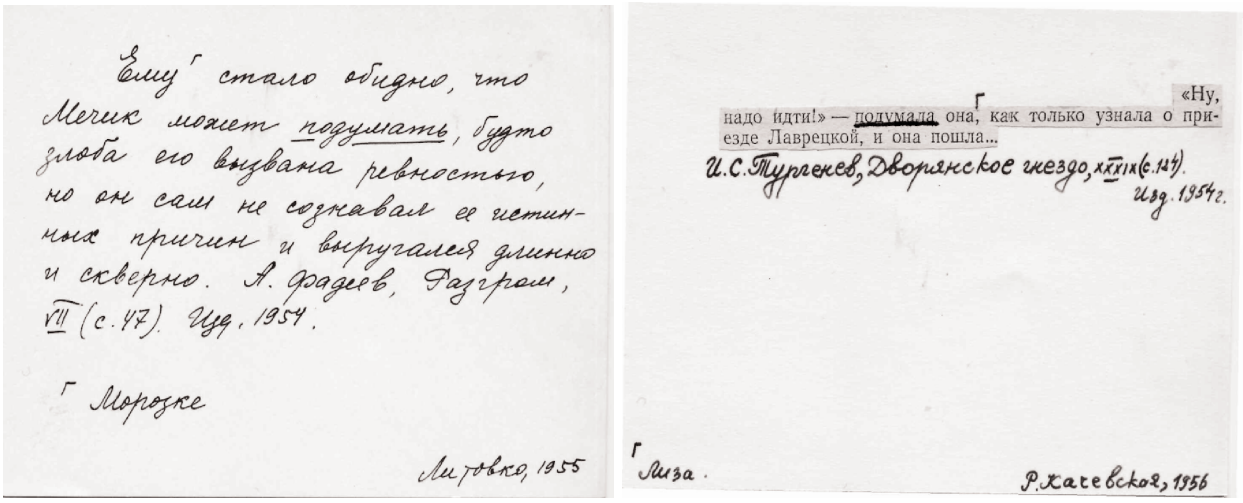

Рис. 1. Образцы словарных карточек

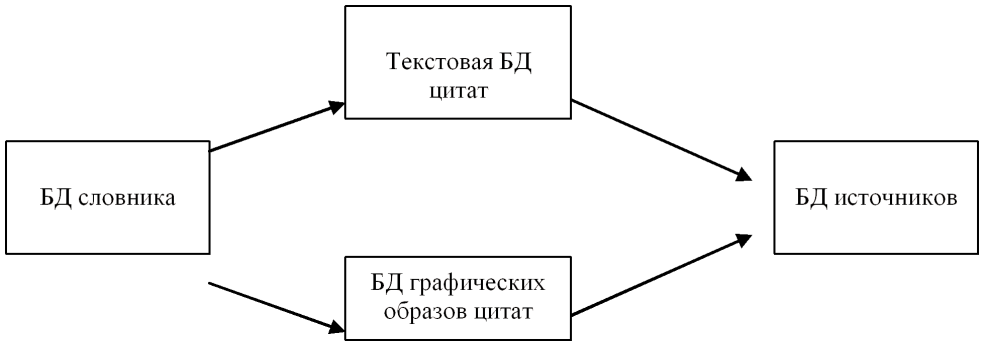

Рис. 2. Информационная модель базы данных БСК

Третий массив – это сводный библиографический перечень всех источников, из которых делались так называемые выборки.

Схематически идеальную информационную модель БСК можно представить в следующем виде: (рис.2)

Первое и главное, что должно быть создано, это база данных словника. В настоящее время не существует даже полного бумажного словника. Наличие электронного словника позволит получить разные статистические данные: количество разных слов, количество карточек в картотеке, количество составных слов, то же в разрезе отдельных букв и т.п. Также появляется возможность сравнивать с ним различные словоуказатели, чтобы выявить, какие слова представлены в БСК, а какие нет.

Формат БД словника имеет следующую структуру:

Слово	Количество карточек	Отсылка	Примечания	Ударение

Если встречается составное слово через дефис (например: *старик-боровик; багряно-красный*), то для него создается две и больше записей соответственно числу составляющих. Например:

багряно-красный	4			
……	……	……	……	……
красный<=		багряно-красный		

В настоящий момент ведется обработка данных карточек новой картотеки и создается база данных словника БСК в СУБД Access. По состоянию на конец 2006. введено около 70 тысяч записей.

Вторая база, которая может и должна быть создана – это библиографическая база источников картотеки, связанная с БД словника. Ее наличие даст возможность получать справки о том, из каких источников отбирались цитаты для того или другого слова. Библиометрический анализ этой базы позволит сказать, к каким жанрам и к каким периодам времени относятся эти источники, какие авторы в каком объеме представлены в картотеке и т.п. Эта информация будет полезна при подборе новых источников и новых цитат.

От словника должны быть организованы ссылки к текстовой базе самих словарных карточек с цитатами. Наличие такой базы позволило бы работать с ней как с корпусом текстов. Но эта задача на сегодня вряд ли выполнима. Карточки-цитаты представляют собой, в основном, рукописный материал, поэтому в большинстве случаев ввод с помощью стандартных процедур сканирования и распознавания с созданием текстового массива невозможен. В этом случае можно рассмотреть альтернативу – хранение в базе электронных графических образов карточек. Но и этот вариант требует огромных трудозатрат и специального оборудования.

Одновременно стоит вопрос о пополнении картотеки. В настоящем виде БСК недостаточно репрезентативна. Это объясняется как ее внутренними изъянами (невключение ряда авторов и материалов в советское время по причинам идеологического характера), так и тем, что уже примерно 15 лет она фактически не пополняется по причине нехватки финансовых средств.

Сегодня главное в формировании и реформировании БСК – это вопрос об источниках и перспективах нового Академического словаря. Совершенно очевидно, что этот словарь должен учесть гораздо большее число типов текстов, чем то, которое содержится в картотеке. И очевидно, что только новые современные технологии (электронные библиотеки, корпуса текстов, программные средства автоматизация труда лексикографов) могут обеспечить необходимый прогресс. Пополнение картотеки должно делаться уже в электронной форме. Таким образом, следующий массив данных – это электронная база данных цитат, дополняющая БСК. При этом следует проработать вопросы связывания двух картотек, бумажной и электронной, и их совместного использования.

Для хранения цитат создана база данных словника в СУБД Microsoft Access, включающая в себя 4 таблицы: **Авторы** (список авторов произведений, из которых берутся цитаты), **Произведения** (список произведений с выходными данными), **Жанры** (проза, поэзия, публицистика и т.п.), **Цитаты** (список цитат, содержащих ключевые слова – аналогично карточкам словарной картотеки) (см. ниже рис. 3).

Цитаты (аналоги словарных карточек в этой базе) являются гиперссылками к полнотекстовой базе данных, где хранятся полные тексты первоисточников (при их наличии). При этом БД основного словника связывается с БД электронных цитат посредством гипертекстовой навигации (язык HTML). При этом сохраняется возможность автономного поиска в каждой отдельной базе.

На втором этапе работы должно быть выбрана интегрированная программная среда, обеспечивающая стыковку указанных выше баз данных с корпусом текстов в формате XML–TEI.

Ключевое слово	Фразема	Цитата-гиперссылка	Примечания	Источник
автокефальный		Местные христиане — римские католики, но в последнее время церковь Катманду активно добивается статуса **АВТОКЕФАЛЬНОЙ**.		Вести из Непала
англичанин		- А вот в севастопольскую кампанию... - начал седой солдат, будто покрытый зеленью от старости -...АНГЛИЧАНЕ к нам приходили, в Соловки палили.../ - Так то АНГЛИЧАНЕ, а не немцы, дедушка... - **АНГЛИЧАНЕ, АНГЛИЧАНЕ**...		Кирикова лодка

Рис.3. Фрагмент таблицы "Цитаты в MS Access"

Следующий вопрос, требующий обсуждения – это выработка принципов «сосуществования» картотеки с корпусами текстов. В последние десятилетия в обиход лингвистов вошел новый инструмент лингвистических исследований, получивший название корпуса текстов. На Западе корпусная лингвистика сформировалась как отдельный раздел науки о языке в первой половине 1990-х годов. В начале 21-го в. корпусная лингвистика стала развиваться и в России. Самый большой российский проект – это Национальной корпус русского языка, созданный в рамках программы «Филология и информатика» РАН (ruscorpora.ru) [Национальный корпус 2005, Герд et al. 2004, Перцов 2006].

Назначение языкового корпуса – показать функционирование лингвистических единиц на большом материале и в их естественном окружении – контекстной среде. Понятие корпуса является продолжением традиционных картотек, с которыми всегда работали лингвисты. Однако картотеки не дают возможности обратиться к более широкому контексту, не позволяют получить сводную статистику об употреблении тех или иных единиц в целом в языке или подъязыке. Характерная особенность современных корпусов – наличие в текстах специальной разметки, метаданных. Набор этих метаданных во многом определяет возможности, предоставляемые корпусом исследователям.

Поисковые возможности корпусов (корпусных менеджеров) намного превосходят возможности картотек. Они включают поиск конкретных словоформ; поиск словоформ по леммам; поиск группы словоформ в виде разрывной или неразрывной синтагмы; поиск словоформ по набору морфологических признаков; отображение информации о происхождении, типе текста и т.п.; вывод результатов поиска с указанием контекста заданной длины; сохранение отобранных строк конкорданса в отдельном файле на компьютере пользователя; и мн. др. [Zakharov 2003]. Использование корпусов позволяет не только изучать лексические единицы в контекстах, но и получать данные о частоте словоформ, частоте лексем, грамматических категорий, о совместной встречаемости лексических единиц (см., напр., [Аверин 2006]), особенностях их сочетаемости, управления и т.д.

Однако национальные корпусы, как правило, содержат лишь морфологическую разметку и мало подходят для нужд практической словарной работы. Представляется, что для решения лексикографических задач требуются метаданные особого рода. Нужна, в частности, лексико-семантическая разметка, фиксация лексико-семантических вариантов слова, когда объектом аннотирования становится не лексема (слово), а отдельное значение. Необходимо также иметь средства тематической и стилистической индексации текстов, которые помогали бы атрибуции лексических единиц. Таким образом, возникает необходимость создания в дополнение к имеющимся картотекам и Национальному корпусу русского языка особого корпуса, который отвечал бы нуждам лексикографии. Его содержательная разметка при этом должна проводиться автоматически, относительно какой-либо заранее установленной системы координат, относительно эталона. Эту систему координат – систему готовых языковых оценок – на данный момент может системно предоставить только какой-либо авторитетный словарь, например, «Словарь современного русского литературного языка». В этих целях, мы полагаем, может быть использован и нормативный «Толковый словарь русского языка» С.И. Ожегова и Н.Ю. Шведовой. Предлагается идея автоматической конверсии существующих толковых словарей в структурированное представление на языке XML. Такие размеченные словари могут использоваться для разметки корпусов. Далее такие корпусы могут использоваться через корпусные менеджеры, а также могут служить источником контекстов русского словоупотребления, специально отобранных лексикографами (картотека цитат в электронном виде), фактически, это будет продолжением и развитием Большой словарной картотеки ИЛИ РАН на базе новых информационных технологий. Представляется, что для решения задач составления словарей уместна вся триада “Интернет – корпус – картотека” [Волков et al. 2004, Захаров 2005 а].

Нам представляется, что в любом случае между корпусом и пользователем-лексикографом должна быть специализированная система, ориентированная на словарные задачи. Очень часто корпусы дают тысячи и десятки тысяч контекстов. Просматривать всё это на экране вряд ли реально как в аспекте времени, так и с точки зрения эффективности. “Нужна некая система тонких семантических фильтров, метаданных, которая поможет найти и отобрать факты для академического словаря, которая превратит грандиозную глыбу, привезённую из

месторождения «РНК», в материал, релевантный для художников и скульпторов от академической лексикографии, ибо она была и остаётся искусством проникновения в семантические глубины отдельных слов" (Герд, 2006, с. 191). Одним из примеров такого фильтра является система так называемых "лексических шаблонов" (word sketches), разрабатываемая английскими и чешскими исследователями [Kilgariff et al. 2004, Rychly et al. 2004, Pala 2006].

В конечном счете – уже за рамками работ по автоматизации Большой словарной картотеки ИЛИ РАН – требуется разработка интегрированной системы (условно "АРМ лексикографа"), которая в качестве составных частей, кроме картотечных баз данных и электронной картотеки цитат, кроме корпуса, включала бы и различные словари, и средства лингвостатистической обработки данных корпуса и/или картотек. Задача такой системы – обеспечить лексикографов необходимым и достаточным лексическим массивом и инструментарием, позволяющими получать объективную информацию о слове, его связях с другими, классифицировать контексты слова и т.п. – в общем, проводить научные изыскания и составление словарных материалов. Короче говоря, необходимо дать пользователю программно-лингвистическую систему, функциональные возможности которой позволят эффективно оперировать данными корпуса в сочетании с картотеками и словарями.

Список литературы

1. Аверин А.Н. Разработка сервиса поиска биграмм // Труды международной конференции «Корпусная лингвистика–2006». – СПб.: Изд-во С.Петерб. ун-та, 2006. С. 5-15.
2. Большой академический словарь русского языка. Т. 1-6. СПб.: 2004-2006 (издание продолжается).
3. Волков С.Св., Захаров В.П. Информационная среда современной лексикографии: корпус текстов и/или электронная картотека? // Сборник трудов VII Всероссийской объединенной конференции "Технологии информационного общества - Интернет и современное общество".(IST/IMS-2004)– СПб.: 2004. С. 5254.
4. Герд А.С. РНК и академическая лексикография // Труды Международной конференции «Корпусная лингвистика – 2006» (10–14 октября 2006 г., С.-Петербург). СПб.: Издательство С.-Петербургского университета, 2006. С. 88-91.
5. Герд А.С., Захаров В.П. Национальный корпус русского языка в свете проблем современной филологии // Труды международной конференции «Корпусная лингвистика–2004»: Сборник докладов. СПб.: 2004. С. 122-130.
6. Захаров В.П. *(Захаров 2005 а)* Веб-пространство как языковой корпус // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог-2005" (Звенигород, 1-6 июня 2005 г.). М.: 2005 С. 166-171.
7. Захаров В.П. *(Захаров 2005 б)* Компьютерная модель Большой словарной картотеки Института лингвистических исследований РАН // Технологии информационного общества - Интернет и современное общество: Труды VIII Всероссийской объединенной конференции (IST/IMS-2005). СПб.: 2005. С. 3334.
8. Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М.: 2005.
9. Перцов Н.В. О роли корпусов в лингвистических исследованиях // Труды международной конференции «Корпусная лингвистика–2006». – СПб.: Изд-во С.Петерб. ун-та, 2006. С. 318-331.
10. Приемышева М.Н. Картотека как источник лексикографической и научной работы. // Acta Linguistica Petropolitana. Труды Института лингвистических исследований / Отв. редактор Н.Н. Казанский. СПб.: Наука, 2003. С. 23-28.
11. Разработка лексики и фразеологии современного русского литературного языка: Пособие по выборкам. Л.: 1972.
12. Рогожникова Р.П. Большой Словарной картотеке 100 лет // Практическая лексикография. 100 лет словарной картотеке. М.: 1989. С. 14-19.
13. Рогожникова Р.П. Сокровищница русского слова. История большой словарной картотеки Института лингвистических исследований РАН / Отв. редактор Н.Н. Казанский. СПб.: Наука, 2003. 106 с.
14. Сводный словарь современной русской лексики. М.: 1991. Т. 1-2.
15. Kilgariff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of EUROLEX-2004.
16. Pala K. Word Sketches and Semantic Roles // Труды международной конференции «Корпусная лингвистика–2006». – СПб.: Изд-во С.Петерб. ун-та, 2006. С. 307-317.
17. Rychly P., Smrz P. Manatee, Bonito and Word Sketches for Czech . Труды международной конференции «Корпусная лингвистика–2004»: Сборник докладов. СПб., 2004. – С. 324-334.
18. Zakharov V. Russian Corpus of the 19th Century // Text, Speech and Dialogue. Proceedings of the 6th International Conference TSD 2003, České Budějovice, Czech Republic, September 2003 / Václav Matoušek, Pavel Mautner (Eds.). – Springer-Verlag, Berlin, Heidelberg, 2003. – P. 146-151. – (Lecture Notes in Artificial Intelligence, 2807).