ЭЛЕКТРОННЫЙ КОРПУС ТЕКСТОВ XIX ВЕКА[1]

Волков С.Св., Герд А.С., Гринбаум О.Н., Захаров В. П., Панков И.П.

В настоящее время разработка и создание различных корпусов текстов стали настоятельной необходимостью для современной прикладной лингвистики. Подготовка разнообразных по характеру, материалу, практической направленности корпусов текстов – а к работам такого рода уже приступили специалисты в разных научных и образовательных центрах России, в том числе и корпуса текстов XIX века, создаваемого на кафедре математической лингвистики Санкт-Петербургского университета, даст возможность значительно расширить круг лингвистических задач, решаемых на основе современных информационных технологий. Представительные массивы языковых данных в электронной форме позволяют получать разнообразные лингвистические и лингво-статистические данные, изучать динамику процессов изменения лексического состава русского языка, проводить анализ изменения лексико-грамматических характеристик и контекстов в различные периоды времени, в разных жанрах и у разных авторов, и т.д. Корпуса служат также источником и инструментом многоаспектных лексикографических работ по подготовке разнообразных исторических и современных словарей.

Целесообразность создания корпусов определяется двумя предпосылками:

1. данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;
2. достаточно большой (репрезентативный) объем корпуса гарантирует типичность данных.

Корпуса текстов образуют базу корпусной лингвистики – лингвистической дисциплины, сформировавшейся в последние два десятилетия на базе электронной вычислительной техники и изучающей построение лингвистических корпусов, способы обработки данных в корпусах и собственно методологию создания и использования корпусов[i].

Главная и определяющая особенность русского языка XIX века – динамика лексико-семантической системы. Одна из задач корпуса – продемонстрировать, что появлялось, распространялось в языке XIX века и что, наоборот, сокращалось и уходило из употребления. Мы считаем, что создание корпуса текстов XIX века будет способствовать, безусловно, более глубокому и подробному описанию истории русского языка, точной хронологической атрибуции лексических и семантических новаций, детализированной демонстрации факторов семантического, стилистического и фразеологического порядка. Приведем здесь принципиальное мнение В.В. Виноградова по этому вопросу: «…история слова должна производить все содержание, всю цепь смысловых приращений, все «метаморфозы». Она стремится раскрыть конкретные условия употребления слова в разные периоды его речевой жизни»[ii].

В ряде случаев необходима и реальна задача создания полного корпуса соответствующего языка или подъязыка. Очевидно, что для языка XIX века такая постановка дела вряд возможна (по крайней мере, на данном этапе). Основой корпуса XIX века будет **репрезентативная** выборка текстов соответствующего периода, включающая: 1) оригинальные художественные тексты на русском языке (проза и поэзия); 2) переводные сочинения; 3) публицистика; 4) научно-популярная литература; 5) письма, дневники, мемуары; 6) деловые, официальные документы. Важнейшая задача – обеспечить изоморфность выборки системе основных жанров и стилей русских текстов XIX века. В корпус электронных текстов предполагается включать источники, отвечающие следующим требованиям:

- изданные на русском языке и в России;

[1] Работа финансируется Минобразования РФ (грант по фундаментальным исследованиям в области естественных и точных наук №2-19/2002)

- для текстов, изданных и переизданных в XX-XXI вв., предпочтение будет отдано наиболее авторитетным современным изданиям, максимально сохраняющим (при современной графике) орфографию и пунктуацию источника, сверенным по рукописям и прижизненным изданиям.

Для работы с данными корпуса текстов XIX века будет разработана подсистема поиска и статистической обработки данных, предоставляющая исследователям инструмент для решения лингвистических задач. Система предусматривает как стандартные для информационно-поисковых систем (ИПС) функции, так и новые специализированные. В частности, предусматривается решение следующих задач поиска:

- поиск конкретных словоформ;
- поиск словоформ по леммам;
- поиск группы словоформ в виде разрывной или неразрывной синтагмы;
- поиск устойчивых словосочетаний и фразеологических единиц по компонентам;
- поиск устойчивых словосочетаний и фразеологических единиц по заданным моделям (например, все прилагательные по существительному в атрибутивных сочетаниях) ;
- поиск словоформ по набору морфологических признаков;
- просмотр найденных словоформ в определенном контексте;
- отображение информации о происхождении и типе текста;
- просмотр морфологических характеристик заданной словоформы в конкордансе;
- отображение леммы;
- вывод результатов поиска с указанием контекста заданной длины;
- получение различных лексико-грамматических статистических данных;
- сохранение отобранных строк конкорданса в отдельном файле;
- прочие задачи.

Выбор средств представления метаданных в рамках настоящего проекта будет вестись с учетом международного опыта: проекта TEI, сложившихся подходов к применению языков SGML/XML, теоретических и практических принципов формирования уже существующих корпусов, в частности, Британского национального корпуса, Чешского национального корпуса и др. Предполагается также, что метаязыковая разметка подготавливаемого корпуса должна в основном соответствовать рекомендациям EAGLES (Expert Advisory Group on Language Engineering Standards)[iii]. Стандартизация способов представления данных создает: а) предпосылки использования стандартных программных средств обработки текстовой информации; б) условия для многоаспектного использования каждого из текстов, включенных в корпус.

Все тексты будут храниться в системе в трех видах, а именно: а) текстовый архив (в исходном виде); б) банк данных (обработанные, размеченные тексты, предположительно, на языке XML); в) в виде базы данных информационно-поисковой системы (тексты в специальном формате, предназначенном для поиска и статистической обработки). Исходные тексты пройдут несколько этапов конвертирования, в ходе которых будет, при необходимости, осуществлена их перекодировка, структурирование текста, удаление нетекстовых и других элементов, например, иллюстраций, таблиц и т.п., и собственно разметка. В результате такого рода трансформаций будет сформирована структура, содержащая признаки, идентифицирующие такие объекты, как файл, документ, том, часть, раздел, глава, страница, абзац (строфа), предложение, отдельная словоформа. Заголовочная часть документа будет содержать библиографические и типологические признаки (автор, название, источник, год издания, тип текста, жанр и т.п. в соответствии с новаторской системой паспортизации текстов корпуса, разрабатываемой в рамках проекта). В составе программно-лингвистического обеспечения корпуса предполагается наличие автоматизированной программы лемматизации и лингвистической разметки текстов для приписывания словоформам лингвистических характеристик, выбранных в соответствии с поставленными задачами и составляющих лексическую основу информационно-поискового языка системы. Результаты лингвистической разметки будут не только конвертироваться в базу данных (индекс) поисковой системы, но и сохраняться на языке структурной разметки, что даст возможность последующей дополнительной обработки ранее введенного материала.

Специализированный характер информационно-поисковой системы по корпусу текстов XIX века определяет ее весьма существенные особенности. Сохраняя все черты и возможности документальных ИПС, по типу релевантности система фактически представляет собой особый тип фактографической системы. В основе понятия

релевантности лежит не свойство смысловой близости между двумя текстовыми сущностями (документ и запрос), а точное совпадение лексико-грамматических признаков. При этом в лексический состав информационно-поискового языка (бестезаурусного типа), включается как полнозначная, так и служебная лексика – это определяется самим назначением разрабатываемой системы. Отличия имеют место и в подсистеме выдачи результатов, где наряду с основным видом данных (конкордансов с набором лингвистических и экстралингвистических метаданных, специализированных конкордансов по запросу пользователя), будут осуществляться различные виды статистической обработки лексико-грамматического материала. Система поиска и выдачи результатов базируется на логической (булевой) модели поиска, предусматривающей построение сложных запросов.

Очевидно, что работа по созданию репрезентативного корпуса текстов XIX в. в рамках настоящего проекта неизбежно будет состоять из нескольких этапов, поэтому данный, первый этап рассматривается нами как создание экспериментальной модели корпуса, объективно отражающей языковое состояние XIX века.

Репрезентативность модели обеспечивается изучением подходов других коллективов разработчиков и выработкой научных критериев отбора текстов разных видов, жанров, периодов и т.п. В результате исследования текстового массива XIX в. будут сформулированы адекватно-пропорциональные соотношения текстов разного рода, в соответствии с которыми будет формироваться корпус текстов в реализуемом проекте.

Сегодня многие важные теоретические и практические вопросы составления корпусов намечены, скажем, пунктирно. Одной из таких проблем при составлении корпуса XIX века являются хронологические границы. Казалось бы, здесь все ясно, традиционно и просто. Однако это далеко не так, и существуют различные мнения и варианты.

Если отталкиваться от лексикографической теории и практики, то нижняя хронологическая граница описания словарного состава XIX века, вероятно, может быть установлена по верхней границе Словаря русского языка XVIII века»[iv], которая соответствует 1803-1805 годам. Верхнюю хронологическую границу можно установить в середине 90-х годов XIX века[v]. Однако существуют и другие взгляды. В частности, В.В. Виноградов в его уже ставшей хрестоматийной статье «Семнадцатитомный академический словарь современного русского языка и его значение для языкознания», опубликованной в 1966-ом году, предлагает "отсчитывать" XIX век от Пушкина. "Можно утверждать, что семнадцатитомный словарь поставил перед нами во всю ширь проблему исторического словаря русского языка национальной эпохи или, вернее, нескольких словарей этого времени: от конца XVII в. – до Пушкинской эпохи, от Пушкина – до 50–60-х годов XIX века, от 60-х годов до конца XIX – начала XX века"[vi].

Представляется, что нижнюю хронологическую границу можно установить, начиная с 1810–1812 гг. Основания для этого есть. Позволим напомнить, что, работая над «Проектом «Словаря русского языка XVIII века», Ю.С. Сорокин и Л.Л. Кутина установили верхнюю границу на уровне 1803-1805 гг., с частичным захватом, некоторых более поздних материалов[vii]. Они совершенно справедливо считали, что XVIII век – это языковой период "не в строгих пределах, а с известным «закруглением», с присоединением нескольких десятилетий предшествующего и последующего столетий, где уже обнаруживаются или еще продолжают сказываться характерные для века черты языкового развития, т.е. период существования русского языка от петровской поры до пушкинского времени".

Также и факторы широкого социокультурного контекста подталкивают нас к такому решению, хотя, заметим, Б.А. Ларин предостерегал от преувеличения значения культурно-исторического подхода при интерпретации языковых процессов. Учреждение министерств, заменивших старые коллегии, присоединение Бессарабии, создание антинаполеоновской европейской коалиции, Отечественная война 1812 г. и вызванное ею глобальное изменение менталитета всего русского общества, русские войска в Европе, введение конституционного правления в Финляндии и Польше, освоение русскими Дальнего Востока и Камчатки – вот самый краткий перечень важнейших событий 10-х – 20-х годов XIX в. Самое важное, что с этим периодом связано начало объективных языковых процессов в изменении словарного состава русского языка, например, расширение фонда отвлеченной лексики – (то, что А. С. Пушкин называл «метафизическим языком») – существительные со значением отвлеченного качества, свойства или степени качества и другие группы), обновление состава «простой» лексики, входящей в литературное употребление.

Верхнюю же границу XIX века предлагается поднять на уровень 1904–1905 годов. Это позволило бы, во-первых, включить в язык XIX века Чехова, а также произведения Л. Н. Толстого после «Исповеди». Во-вторых, отразить начавшееся с конца 70-х годов быстрое пополнение лексико-фразеологического фонда, семантическую перестройку многих слов, сопровождающуюся их стилистическим перемещением и определяющую новые соотношения книжного и разговорного пластов в разных стилях, о чем достаточно подробно написал Ю.А. Бельчиков[viii].

В третьих, и это важно, можно будет «закрыть» семантическую историю слов и словообразовательных гнезд, формирование которых преимущественно происходило в языке XIX века. Так, например, слово "алкоголь" вошло в русский язык в самом конце XVIII века (первая лексикографическая фиксация в «Новом словотолкователе» Яновского), большинство его производных (алкогольный, алкоголический, алкоголик, алкоголист, алкоголизм) появляются в XIX веке, а вот слова "алкоголичка" и редкое "алкоголеникотинец" (в системе воззрений Л.Н.

Толстого – человек, не отказавшийся от употребления спиртных напитков и курения) – фиксируются только в материалах начала века.

Кроме того, в результате такого расширения границ, может быть, некоторые новации XX века окажутся, как говорил Б.А. Ларин, «не очень новыми новшествами» – например, аббревиатуры, которые вовсе не «большевистский волапюк», как их называл Дмитрий Мережковский, а вполне системное явление конца XIX – начала XX веков, связанное, по-видимому, с массовым распространением телеграфа.

Не менее важным, чем хронологические границы, является вопрос о жанрах и источниках корпуса. Русский язык XIX века – это огромный массив самых разнородных по происхождению и семантической/стилистической приуроченности слов и словосочетаний: это вся, без ограничений, обширная неология XIX века (как исконная, так и заимствованная), некоторые имена собственные, топонимы (например, устаревшие географические наименования, употребляемые в стилистических целях: Гиперборейское море, Авзония), производные от них, книжно-славянская лексика и в особенности та, которая выполняла специфическую идеологическую функцию в текстах того времени, областные слова, вошедшие в литературное употребление, терминологическая лексика (преимущественно термины, встречающиеся в нескольких терминологических системах или термины, получившие расширительные, переносные, образные значения – ср. инерция, ископаемый, эмбрион и т.п.), лексика профессиональных арго. Какова должна быть доля каждого жанра в общем массиве корпуса? В частности, каково место художественной литературы в языке XIX века? Корпуса современных языков (британский, чешский и др.) обычно отдают художественным произведениям 15-20%. Нам представляется, что в XIX веке этот процент может быть увеличен до 30-40.

Следующая проблема – метатекстовая информация, обеспечивающая многофункциональное использование корпуса различными исследователями. Последняя должна предусматривать, в числе прочего, отлаженную систему паспортизации текстов и систему их тематической и стилистической индексации для атрибуции лексических единиц (ср. пометы в словарях: у рыбаков, у иконописцев, у торговцев, у охотников, у типографов (затычка), у чертежников, у торговцев лошадьми (осечься, понести), в воровском языке, в языке шулеров, в языке чиновников).

Каждый отдельный текст в корпусе, как предполагают его разработчики, будет сопровождаться особым формуляром текста, содержащим основные библиографические и типологические признаки. Формуляр текста представляет собой метаописание текста (метаданные), в соответствии со следующими параметрами:

- полное и краткое наименование текста;
- графика текста (книжно-славянская, современная; русская орфография до 1917 года, использование латинской, греческой и иных видов графики); наличие иноязычных вкраплений в текст; наличие формул, рисунков и т.п.
- атрибуция текста его создателем; атрибуция текста в соответствии с предполагаемой типологией текстов (см., например, типология Б.В. Томашевского);
- фамилия, имя, отчество автора (авторов) текста;
- литературный псевдоним (псевдонимы) автора;
- дата рождения (дата смерти автора);
- датировка создания/датировка первой публикации текста.
- указание на первую публикацию текста (собрание сочинений, отдельное издание, научное издание, журнал, газета, стенограмма, архивные материалы и пр.);
- адрес текста в эталонном научном собрании (издании) сочинений автора с указанием тома и пагинации;
- указание на наличие авторских (редакторских) комментариев, примечаний, исправлений; специальные замечания о подготовке текста к изданию;
- указание на цензора текста (для текстов XIX – начала XX века - в том случае, если это релевантно по культурно-историческим причинам);
- указание на наличие научных (литературных, исторических) комментариев к тексту;
- отсылки к литературно-критическим произведениям, материалам публицистики, другим упоминаниям и аллюзиям данного текста в корпусе текстов; указания на рецензии;
- указание на способ введения текста в корпус (ручной ввод; CD-версии; сканирование; Интернет; оригинал-макет издательства; авторская копия);

- указание лиц, осуществлявших ввод текста в корпус, разметку текста и филологическую проверку текста; указание лиц, подготовивших формуляр текста;
- описание этапов работы над текстом (журнал текста) с указанием данных конкретных исполнителей:
- указание на место хранения текста (указание на учреждение, куда передан текст после обработки – адрес текста).

Есть и много других проблем, которые должны быть решены на первом этапе. Каков круг источников словаря в целом и как правильно определить источники при описании, в частности, научной и терминологической лексики. Должны ли мы привлекать в качестве источников, например, энциклопедические словари XIX века? Сохранять ли в электронных текстах элементы форматирования?

Важнейшая часть работы по созданию корпуса – это грамматическая разметка. Этой теме посвящен отдельный доклад. Здесь подчеркнем лишь то, что корпус создается как морфологически размеченный корпус с возможностью последующего синтаксического анализа. Также считаем крайне важным обеспечить, по возможности, совместимость разметки с Большим корпусом современного русского языка, создаваемым в России в настоящее время[ix].

---

[i] Leech G. The State of Art in Corpus Linguistics // English Corpus Linguistics / Aimer K., Altenberg K.(eds.) – London, 1991. – P. 8-29.

[ii] Виноградов В.В. Чтение древнерусского текста и историко-этимологические каламбуры. ВЯ, 1968, № 1, С. 19.

[iii] www.ilc.pi.cnr.it/EAGLES/home.html

[iv] Словарь русского языка XVIII века. Проект. Л., 1977, С. 10.

[v] Сорокин Ю.С. Основные принципы и источники исторического словаря русского литературного языка XIX века. // Очерки по исторической лексикологии русского языка. – СПб, 1999, С. 39.

[vi] Виноградов В.В. Семнадцатитомный академический словарь русского литературного языка и его значение для советского языкознания. // Вопросы языкознания, 1966, № 6, С. 20.

[vii] М. М. Херасков, Н.Ф. Остолопов, Н.М. Карамзин в период «Вестника Европы», Ив. Ив. Дмитриев, кн. П.И. Шаликов, Ив. Петр. Пнин и, конечно, Г. Р. Державин – вышедшее в 1808 году его Собрание сочинений – своеобразный итог развития русской поэзии XVIII в.

[viii] Бельчиков Ю.А. Некоторые вопросы развития русской разговорной и книжной лексики во второй половине XIX века. Филологические науки, 1975, № 6. С. 185.

[ix] См. http://bokrcorpora.narod.ru