# Экспертная система «Русский текст XIX века»

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

В докладе рассматриваются проблемы создания корпуса текстов – экспертной системы «Русский текст XIX века», обосновывается необходимость ее разработки для современных лингвистических и культурологических исследований, демонстрируется возможность многоаспектного ее использования в разных отраслях знания, принципы конструирования, описываются структурные компоненты экспертной системы. Эмпирическую базу экспертной системы составит репрезентативная выборка из русских текстов XIX века различных жанров: художественной литературы, публицистики, научных текстов, деловых документов и пр. Особое внимание в докладе уделяется вопросам обработки текстов перед включением в корпус, снятию лингвистических трудностей, проблемам морфологической разметки текстов и введения метаязыковой информации.

Система представляет собой комплекс лингвистических, информационных и программных средств для исследователей историков русского языка и обеспечивает получение различных данных для решения фундаментальных и прикладных филологических задач.
Сейчас в обиход лингвистов начинает входить класс систем, получивший название „корпус текстов“. Назначение языкового корпуса – показать функционирование лингвистических единиц на большом материале и в их естественном окружении - контекстной среде. Основная особенность языковых корпусов – наличие в составе текстов специальных идентификаторов, характеристик, описывающих единицы текста, относящиеся к различным языковым уровням. Размеченные корпуса, т.е. корпуса, в которых все интегрированные в них языковые единицы (а в составе корпуса – уже не языковые, а корпусные единицы, это уже новая языковая действительность) получили строго заданный создателями корпуса набор признаков, могут использоваться для получения разнообразных лингвистических данных. В частности, на их основе можно просмотреть искомые единицы в различных контекстах, получить данные о частоте словоформ, лексем, грамматических категорий, о совместной встречаемости лексических единиц, особенностях их сочетаемости, управления и т.д. Данные корпусов могут быть использованы в целях обучения как родному, так и неродному языку. Корпуса служат также источником и инструментом многоаспектных лексикографических работ по подготовке разнообразных словарей.
Возможности такой системы зависят, в первую очередь, от типа разметки. Создаваемая в Санкт-Петербургском университете система "Русский текст XIX века" в качестве важнейшей составной части включает в себя корпус текстов XIX века с морфологической разметкой. Кроме того, в ее состав входят и другие компоненты, необходимые современному филологу. Многоаспектное использование системы обеспечивается введением в нее разнообразной метатекстовой и лингвистической информации. Метаописание текстов, помимо данных о конкретном тексте – время создания, автор, источник, процессинг текста, в ряде случаев

включает специальный лингвистический комментарий (характеристика устаревших слов и форм, индивидуально-авторских образований, неологизмов, фразеологизмов и устойчивых словосочетаний, заимствованных слов и экзотизмов, диалектизмов, профессиональной лексики и под.).

Основой корпуса текстов является репрезентативная выборка текстов, обеспечивающая изоморфность текстового массива корпуса реальному массиву текстов XIX столетия. Устанавливаются два основных класса текстов: 1) прагматичные тексты (faction), содержащие, в основном, фактическую, «сущностную», деловую информацию и 2) беллетристические тексты (fiction). Внутри этих классов выделяются несколько разнообразных по жанрово-тематической принадлежности и особенностям содержания, композиции, структуры и стиля видов текстов, в том числе: 1) оригинальные и переводные художественные тексты на русском языке (проза, драматургия, поэзия); 2) публицистика и литературная критика; 3) научные и научно-популярные сочинения, научные и научно-популярные журналы; 4) письма, дневники, мемуары, записки; 5) деловые, официальные, политические и правовые документы, законодательные акты; 6) религиозная литература; 7) фольклорные произведения, 8) этнографические тексты, описания народного быта; 9) литература по военному, морскому делу; 10) географические и экономические описания, тексты по сельскому хозяйству, земледелию, промышленности; 11) тексты, отражающие профессиональную речь и арго и нек. др.

Работа по созданию корпуса проводится в несколько этапов:

***1. Предобработка текста (подготовительный этап)***

На этом этапе принимается решение о включении текста в корпус, проводится его лингвистическая экспертиза, осуществляется подготовка библиографического описания текста. Все тексты проходят филологическую выверку. Затем осуществляется удаление или преобразование нетекстовых элементов (рисунки, таблицы, формулы и т.п.).

Для текстов 19 века сложным является вопрос о сохранении исконной графики текста. На данный момент принято решение об использовании в корпусе средств современной графики. Отчасти это обусловлено тем, что многие тексты для сканирования берутся по современным переизданиям.

В то же время для некоторых исследователей представляла бы интерес возможность учета старой орфографии. Например, как известно, буквы «фита» и «ижица» служили до реформы орфографии 1917-го года устойчивым графическим признаком заимствованных на разных этапах истории русского языка греческих слов, и их наличие было бы, возможно, полезно для исследователей. Буква ѳ релевантна для некоторых грамматических форм. Поэтому предполагается небольшую часть корпуса сделать со старой орфографией.

Разнообразная лексическая вариантность, отличающая тексты 19 века от современной нормы на этапе подготовки текста не устраняется (ср*., анатом и анатомист, аттака и атака, басурман – бусурман, бархотный и бархатный, бедность* и *бедство*, *ветер* и *ветр, древлий и древний, виолончель* и *виолончеля, квартира* и *квартера, мания* и *маниа*, *волкан и вулкан*, *мародер* и *мародера, найти* и *найтить, онбар* и *амбар, пугать* и *пужать, самодовольствие – самодовольство – самодоволие, тяхчае* и *тяжелее* и др.). Это же относится и к варьированию, обусловленному исторически сложившимся взаимодействием книжнославянской и русской стихий (ср., например, варьирование окончаний *-ыя(-ия)* / *-ой (-ей)* в род. падеже женского рода имен прилагательных: «Когда под скипетром великия жены // Венчалась славою счастливая Россия» (*Пушкин А.С.* «Воспоминания в Царском Селе»[1]), ср. также характерную для XIX века форму Творительного падежа ед. числа *благодатию* и

[1] *Пушкин А.С.* Стихотворения 1813-1820 гг. Полн. собр. соч. в 10 томах. М., 1957. Т. 1. С. 83.

современную *благодатью*. Очевидно, что в данном случае вариативность должна быть сохранена, так как именно этот аспект может быть предметом лингвистического исследования.

### *2. Разметка текста*

Осуществляются три типа разметки, которым соответствуют три набора метаданных: содержательно-библиографические, структурно-графические и лингвистические.
Выбор средств представления метаданных ведется с учетом международных стандартов и рекомендаций (проект TEI (Text Encoding Initiative)[2], EAGLES (Expert Advisory Group on Language Engineering Standards)[3], ISLE (International Standards Language Engineering)[4] и др.)
Первый уровень метаописания текстов корпуса включает набор стандартных библиографических элементов данных и набор признаков, характеризующих жанровые и стилевые особенности текста.
Второй уровень метаописания текстов, входящих в состав корпуса, включает набор формальных и структурных признаков текстов. К числу формальных признаков относятся: имя файла, сопровождаемое данными об истории создания и обработки данного экземпляра документа, параметрах кодирования, версии языка разметки, исполнителях работ (этапов работ). Структура документа включает следующие элементы: текст, раздел (глава), абзац, предложение, слово.
Третий уровень разметки – собственно лингвистическая разметка, заключающаяся в лемматизации (восстановлении нормативной формы словоформ) и приписывании всем словоформам леммы и морфологических характеристик. Для разметки используется система „Диалинг“[5]. Результаты выдаются в виде размеченного файла на языке XML. Результаты автоматической разметки далее проверяются, корректируются и дополняются вручную.
Результаты анализа размеченных текстов позволили выделить разнообразные проблемы, требующие решения. В их числе следует назвать:

- анализ слов с дефисом. Программа разбивает такие слова на два, что часто является ошибкой (*кто-нибудь, бакш-таги*). В других же случаях дефис действительно разделяет два слова (*старик-охотник*). Возможны случаи, когда написание ряда слов и словосочетаний через дефис было традиционно для орфографии XIX века или предписывалось орфографическими нормами этого времени (*Бог-знает*; *так-называемый*; *на-лету*; *быть-может и пр.*)[6];
- нераспознавание многих имен собственных;
- нераспознавание прилагательных и существительных, образованных от имен собственных;
- аббревиатуры;
- сложные слова различных типов (см. разд. 4);
- словообразовательные дериваты и др.

Важной особенностью разрабатываемой системы является учет особенностей языка XIX века (графическая, морфологическая, фономорфологическая, словообразовательная и лексическая вариантность, конкуренция норм и пр.). По мере накопления материала модуль

---

[2] *Sperberg-McQueen, C. M., Burnard, L.* (eds.). Guidelines for Electronic Text Encoding and Interchange. (2001) URL: http://www.hcu.ox.ac.uk/TEI/P4X/index.html, также *Recommendations for the Morphosyntactic Annotation of Corpora*, EAG-TCWG-MAC/R. См. ftp://ftp.ilc.pi.cnr.it/pub/eagles/corpora/annotate.ps.gz

[3] См. http://www.ilc.pi.cnr.it/EAGLES96/browse.html

[4] Calzolari N. et al. The ISLE in the Ocean. Transatlantic Standards for Multilingual Lexicon.

[5] http://www.aot.ru

[6] Авторы благодарят студентку II курса отделения математической лингвистики филологического факультета СПбГУ Хохлову М.В., предоставившую в их распоряжение подробный список такого рода примеров.

морфологического анализа и разметки должен быть адаптирован к языку 19 века. В частности, можно кратко перечислить следующие проблемы, связанные с особенностями лексики и морфологии 19 века:

- фонетические варианты (3.3.1);
- словообразовательные варианты и дублеты (3.4);
- устаревшие и редкие слова (3.5);
- устаревшие морфологические формы (3.6) и др.

В докладе приводятся также параметры описания текста как целого и его структурных компонентов, типология текстов, технология подготовки текстов для системы, критерии отбора источников и/или их фрагментов из электронной библиотеки в корпус, принципы и метаданные лингвистической разметки текста.

В настоящее время у нас нет настоящего словаря русского языка XIX века[7], отвечающего задачам современных исследований. И безусловно, данная система в целом и отдельные словники, которые будут созданы на ее основе, смогут восполнить его отсутствие. Однако словарь, даже при его наличии, всегда представляет лишь основную номенклатуру лексических единиц, как правило, одиночных слов в их основных значениях. На самом деле корпус должен рассматриваться как неотъемлемое дополнение к словарю. Словари и грамматики почти не учитывают неустойчивую и вероятностную природу языка. Существует огромное число лексико-семантических вариантов слова, лексических единиц и устойчивых словосочетаний, которые никогда не попадут в словарь (так, например, редкое сейчас слово **дагерротип** ‘*фотографическое изображение на металлических пластинах, выполненное с помощью метода, предложенного французским изобретателем Луи Дагерром*’ в середине XIX века активно использовалось передовой русской критикой и публицистикой для неодобрительного обозначения буквального копирования действительности в литературе, отсутствия художественного обобщения: «Обратите внимание на современную литературу и критику. .. Дагерротипы она выдает за картины и тратит свои силы на отыскивание красот в мелочах» (Н.В. Шелгунов); значение это не фиксируется даже самым полным сейчас «Словарем современного русского литературного языка» в 17-ти томах), корпус же предоставляет средства для их выявления и учета. Также лишь частично учитывается в словарях синонимия. Корпус позволяет просмотреть необходимые единицы в их естественных контекстах и выявить различные оттенки значения. То же следует сказать об изучении лексической сочетаемости, где корпус предоставляет благодатный материал. Существует элементы текста, которые без корпуса практически не возможно сколь-либо серьезно изучать, например, разрывные словосочетания или вкрапления в русский текст слов на иностранных языках. И так далее.

Можно сказать, что только сейчас лингвистика получает возможность стать точной эмпирической наукой. До сих пор сведений для полного и точного описания языка всегда не хватало, хотя этот факт и не был, возможно, осознан и явно сформулирован. Можно сказать также, что только на основе можно приступить к описанию лексической и грамматической семантики языка. Корпус XIX века интересен еще и тем, что позволит проводить сравнительные исследования, наблюдая развитие русского языка в XX и XXI веке.

---

[7] См. об этом *Сорокин Ю.С.* Основные принципы и источники исторического словаря русского литературного языка XIX века. // Очерки по исторической лексикологии русского языка. Памяти Ю.С. Сорокина. СПб., 1999, С. 29 – 39; *Волков С.Св.* Об «Историческом словаре русского языка XIX века». // Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН. Том I. Часть 3. СПб, 2003, С. 85-94.

Для извлечения из корпуса новой, ранее не доступной информации необходима разработка как соответствующих средств, так и методов, что будет делаться в процессе опытной эксплуатации экспертной системы.
Уже говорилось, что корпус является важнейшей, но не единственной частью системы. В нее также входят: 1) электронная библиотека с возможностью навигации и поиска; 3) комплекс средств, обеспечивающих проведение лингвистами самостоятельной обработки языкового материала, представленного в электронной библиотеке; 4) библиография работ учебного и научного характера, посвященных русскому языку 19 века, с последующим переходом к хрестоматии таких работ; 5) различного рода словники и глоссарии, в том числе созданные на материале корпуса.
Электронная библиотека представляет собой филологически выверенные полные тексты различных видов и жанров. В целях обеспечения репрезентативности каждый вид в корпусе представлен небольшими текстами или фрагментами. Поэтому полные тексты хранятся отдельно с сохранением всех типографских особенностей конкретного издания.
Пользователю будет предоставлен набор процедур и программных средств для их обработки. Это позволит получить данные не только по языку XIX века в целом, но и, скажем, по языку отдельного писателя. Например, можно получить характеристики стиля, проследить изменение частот и контекстов для определенных слов у какого-либо автора в различные периоды времени. Таким образом, система будет интересна не только лингвистам, но и литературоведам.