(КШС), состоящая из взаимосвязанных системы кодирования символов и семейства шрифтов для их отображения. Преимущество КШС состоит в том, что в каждом шрифте одного семейства один и тот же символ, независимо от своих преобразований и особенностей, имеет один и тот же код, что значительно облегчает задачи обработки и конвертирования текстов.

Одна из задач при создании и развитии КШС состоит в том, чтобы правильно классифицировать новый буквенный или небуквенный символ, отнести его к определенной гарнитуре и шрифту и расположить на соответствующем коде.

Механизм внесения изменений в КШС «Манускрипт» заключается в изменении набора символов базы данных, семейства шрифтов и документации по КШС.

Приведение набора символов базы данных информационно-поисковой системы «Манускрипт» к многобайтовому набору символов UTF8LAPREXT1, поддерживающему концепцию КШС, привело к некоторым техническим проблемам, в частности:

– к необходимости уменьшения длины нелатинских (кириллических) имен объектов базы данных до 15 символов (следствие многобайтовости и ограничений выбранной СУБД),

– к потребности в лингвистической сортировке символов.

Лингвистическая сортировка, в отличие от бинарной, позволяет сортировать символы в соответствии с их алфавитным порядком, а не их числовым представлением (кодом) в КШС. Использование лингвистической сортировки, которая может иметь две разновидности — одноязыковую и многоязыковую, обусловлено также необходимостью подготовки перечней текстовых единиц с порядком следования, задаваемым пользователем.



# Корпус русского языка XVIII века: текущее состояние[1]

В. Д. Соловьев, Р. Б. Ахтямов
Казанский государственный университет, Россия

*The paper is devoted to the main tasks and the intermediate results of the project aimed at creation of corpora of the XVIII century Russian language. The main achievement is digital representation of the great number of books and journals and the solution of the image clearance problem. The perspectives of development are discussed.*

Восемнадцатый век представляет собой один из наиболее интересных периодов развития русского языка. В это время произошло резкое изменение языка — от древнерусского к современному. Вместе с тем русский язык XVIII века явно недостаточно изучен, мало внимания ему уделяли и компьютерные лингвисты. До настоящего времени не были созданы электронные словари, в Национальном корпусе русского языка [Национальный 2006] XVIII век представлен лишь несколькими источниками, в основном, Карамзиным. Большее количество текстов можно найти в Интернете (обзор см. в [Русский 2003: 123–131]). Однако и там основное внимание уделяется одному автору — Ломоносову. Кроме того, подавляющее большинство ресурсов представляют собой осовремененные тексты, приведенные в современной орфографии.

В совместном научно-исследовательском Центре «Культурное наследие и информационные технологии» Казанского государственного университета и АН Республики Татарстан работа по созданию корпуса русского языка XVIII века ведется с 2002 г. [Исследования 2002: 21–26]. На первом этапе (к 2006 г.) была создана электронная библиотека. Она создавалась на базе Научной библиотеки КГУ и библиотеки Казанского научного центра РАН, ко-

[1] Работа выполнена при поддержке РФФИ, грант № 04-06-80050, и РГНФ, грант № 04-04-12042в.

торые обладают богатыми фондами книг XVIII века. Для сканирования применялись сначала обычные сканеры и цифровой фотоаппарат, а затем сберегающий старинную бумагу и высокопроизводительный планетарный сканер ЭЛАП ПЛАН СКАН. При сканировании было выбрано разрешение 300 dpi, глубина цвета — 24 бита, графический формат — tiff.

К настоящему времени отсканировано около 25 тыс. страниц. Часть отсканированных материалов доступна через Интернет [Электронная 2006], остальные материалы размещены в локальной сети библиотеки КГУ и доступны в компьютерных классах. По адресу http://lsl.ksu.ru/i835.htm помещена информация о созданных ресурсах. Для представления графических файлов в Интернете выбран ориентированный на представление старинных книг и рукописей формат DjVu и система DjVu Editor 3.5 — недавняя разработка фирмы Lizard Tech Inc., находящаяся в свободном доступе [Электронная 2006].

Следующая задача, к решению которой мы приступили, — распознавание текстов XVIII века с сохранением оригинальной орфографии. Существующие системы распознавания, вроде Fine Reader, не справляются с этой задачей. Причины две: плохая сохранность оригинала и отсутствие словарей, которые можно было бы подключить к системе и использовать для проверки распознаваемого слова.

Мы начали с решения второй задачи. К настоящему времени создан электронный словарь на базе Словаря Академии Российской. Для кодировки символов старой русской орфографии выбран Unicode и шрифт Palatino Linotype, ставшие de facto стандартами.

Затем была создана система работы с текстами и словарями XVIII века (автоматизированное рабочее место лингвиста-лексикографа), позволяющая выполнять ряд операций, в том числе — перевод текстов в современную орфографию, построение по заданному тексту словаря и объединение словарей, что позволяет расширять исходный словарь. АРМ лингвиста содержит также средства сортировки [Исследования 2005: 133–138], что позволило создать обратный словарь XVIII века.

Для решения первой проблемы — эффективной обработки изображений низкого качества — создана система реставрации [Электронные 2005: 249–253], позволяющая существенно повы-





сить качество изображений. Разработанная система обеспечивает более высокое качество реставрации по сравнению с известными зарубежными аналогами [EVA 2005: 186–191]. На конференции EVA ей посвящен отдельный доклад. В связи с трудностями интеграции в Fine Reader (он не позволяет подключить пользовательский словарь старорусского языка) в дальнейшем планируется создание собственной системы распознавания символов, ориентированной на русский язык XVIII века.

Наконец, конечной целью работ является создание Большого корпуса русского языка XVIII века. Предполагается создание двух его версий — в оригинальной орфографии и в современной.

Корпус на основе современной орфографии строится с привлечением текстов, размещенных в Интернете, в первую очередь в Русской виртуальной библиотеке [Русская 2006] и на сайте немецкого издательства ImVerden [ImVerden 2006]. Достигнута договоренность о включении этого корпуса в состав Национального корпуса. Начата его метаразметка в соответствии с технологией Национального корпуса [Национальный 2005: 62–88], к настоящему времени размечен двухтомник Фонвизина из Русской виртуальной библиотеки.

Корпус в оригинальной орфографии создается в основном на базе отсканированных нами книг. Часть имеющихся графических файлов была переведена с помощью Fine Reader в текстовый формат с последующим постредактированием. Для адаптации системы FineReader к текстам XVIII века разработаны шаблоны всех букв старорусского алфавита (пользовательские шаблоны) и осуществлена настройка языка «Русский (старая орфография)». В режиме обучения системы осуществлялось наращивание базы шаблонов символов на основе 30 страниц из книги «Древняя Российская вивлиофика», часть 10. Экспериментально установлено, что этого объема достаточно для построения эффективной базы шаблонов. После настройки языка с подключением встроенного и пользовательского шаблонов удалось добиться точности распознавания текста около 90%.

Следующей проблемой является морфологическая разметка корпуса. Для этой цели адаптирован морфологический анализатор, ранее созданный в Лаборатории компьютерной лексикографии и лексикологии МГУ [Труды 2006]. Наибольшие сложности возни-

кают при обработке текстов первой трети XVIII века. Это связано с большой вариативностью языковых единиц. Текстам этого периода свойственна слабая степень кодификации и конкуренция средств, принадлежавших в предшествующий исторический период к разным стилям и подъязыкам. В порядке эксперимента лематизатор применен к произведениям Кантемира. Получен прямой, обратный и частотный словари Кантемира. С помощью программы «Диктум», созданной в той же лаборатории, построен конкорданс словоформ. Далее планируется уточнение характеристик единиц корпуса (начальные формы, грамматические значения), а также существенное увеличение объема корпуса.

### Список литературы

Кукушкина и др. 2006 — *Кукушкина, О. В.* Вариативность языковых единиц в корпусе русских текстов первой трети XVIII в. и проблема автоматизации его морфологического анализа : тр. школы по компьютерной лингвистике / О. В. Кукушкина, А. А Поликарпов. — Казань : Отечество, 2006. (В печати).

Маргулис 2002 — *Маргулис, И. С.* Раскладка клавиатуры для работы с русскоязычными текстами XVIII–XIX веков / И. С. Маргулис // Исследования по информатике. — № 4. — Казань : Отечество, 2002.

Национальный 2006 — Национальный корпус русского языка [Электронный ресурс]. — 2006. — Режим доступа: http://www.ruscorpora.ru, свободный. — Загл. с экрана.

Национальный 2005 — Национальный корпус русского языка: 2003–2005. — М. : Индрик, 2005.

Русская 2006 — Русская виртуальная библиотека [Электронный ресурс]. — 2006. — Режим доступа: http://www.ruscorpora.ru, свободный. — Загл. с экрана.

Скоробогатов 2003 — *Скоробогатов, А. В.* Письменная культура России XVIII века в сети Рунет / А. В. Скоробогатов // Русский язык в Интернете / ред. В. Д. Соловьев — Казань : Отечество, 2003.

Соловьев и др. 2002 — *Соловьев, В. Д.* Электронная коллекция древних книг и рукописей / В. Д. Соловьев, Г. А. Николаев // Исследования по информатике. — № 4. — Казань : Отечество, 2002.

Электронная 2006 — Электронная библиотека книг XVIII века [Электронный ресурс]. — 2006. — Режим доступа: http://it.knc.ru/book, свободный. — Загл. с экрана.

Южиков 2005 — *Южиков, В. С.* Автоматизированная система реставрации и обработки изображений старопечатных текстов и рукописей / В. С. Южиков // Электронные библиотеки : Перспективные методы и технологии, электронные коллекции : тр. 7-ой Всерос. конф. — Ярославль : Яросл. госун-т, 2005.

ImVerden 2006 — ImVerden. URL: http://www.imwerden.de. 2006

Stanco et al. 2005 — Stanco F., Ramponi G., Russo W., Pelusi S., Mauro P. Digital automates restoration of manuscripts and antique printed books. EVA 2005 Florence. Proceedings. Bologna, 2005.