Кодировочно-шрифтовая система «Манускрипт»: классификация символов и технологические вопросы их представления в базе данных

В. А. Романенко
Удмуртский государственный университет, Ижевск, Россия

The classification of symbol types in ancient Slavonic manuscripts on base symbol, functional variant and inscription variant is presented as well as its implementation in information-retrieval system "Manuscript" as character-font encoding scheme (CFES). The mechanism of updating the CFES with new symbols is described. Some important questions of storage and linguistic sort of multi-byte symbols in the database are discussed.

В настоящее время одной из актуальных задач при подготовке публикаций древних славянских рукописей является необходимость как можно более точного воссоздания всех особенностей шрифта, оформления и декоративных элементов в компьютерном наборе текстов. Актуальность обусловлена тем, что при подготовке транскрипции текста для печатного или электронного издания невозможно сохранить все исходное многообразие начертаний символов в связи с фиксированным их набором и с отсутствием в существующих компьютерных шрифтах вариантов представления конкретного символа.

К сожалению, стандартная система кодирования символов Unicode не содержит даже полного перечня основных символов кириллического алфавита. Поэтому типичный подход к полному сохранению всех существенных особенностей начертания символов, используемый во многих проектах (Titus, Sofia — Trondheim, Манускрипт и др.) состоит в максимально возможном использовании стандартного диапазона Unicode и в расширении его недостающими символами и вариантами начертания. В результате в каждом проекте (а порой и для каждого конкретного текста!) создаются свои системы кодирования символов и шрифтовые гарнитуры, что создает колоссальные трудности при пользовании электрон-

ными коллекциями и библиотеками, при обмене данными, при автоматизированной обработке текстов и при объединении результатов работы разных коллективов в совместные электронные коллекции и библиотеки.

В данном докладе представлена классификация типов символов, позволяющая провести границу между базовой и вариативной составляющими начертания символов одним писцом, разделить при компьютерной обработке почерки разных писцов и при этом отобразить особенности начертания символов. В основе классификации символов лежит деление их на основные символы, их функциональные варианты и варианты начертания.

1) Основные символы и их функциональные варианты (например, ѥ — є, ⰾ — ⰾ҃) представлены отдельным знакоместом во всех шрифтах и шрифтовых гарнитурах.

2) Вариант начертания — вариант основного символа (или функционального варианта), характеризующийся особенностями, могущими иметь значение при анализе рукописи. Каждый вариант начертания связан с базовым символом и классом преобразования, которому подвергнут основной символ.

Примеры классов преобразований:

– геометрические движения (сдвиги вверх, вниз, влево, вправо),

– геометрические отражения (вертикальное, горизонтальное),

– геометрические деформации (уменьшение, увеличение, удлинение, наклон),

– изменение веса (жирный, двойной, тройной),

– дополнение элементами (наличие/отсутствие перекладин, перечеркивания, наличие точки, креста и других элементов).

3) Почерки разных писцов, значимые для исследователей рукописи, отражаются различными гарнитурами шрифтов (гарнитуры Menaion и Panteleymon — для представления почерков основных писцов Путятиной минеи XI века и Пантелеймонова Евангелия рубежа XII–XIII веков.

Таким образом, каждое индивидуальное начертание символа характеризуется тремя составляющими: кодом символа (знакоместом), типом преобразования (шрифт) и типом почерка (шрифтовая гарнитура).

Эта классификация реализована в информационно-поисковой системе «Манускрипт» как кодировочно-шрифтовая система

(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.