# К ВОПРОСУ О ПРЕДСТАВЛЕНИИ МНОГОЯЗЫЧНЫХ ТЕКСТОВ С ДИАКРИТИЧЕСКИМИ ЗНАКАМИ*

Д.Е. Кондратьев, О.В. Тихонова
Центр «Проблемы исторического познания»
Институт всеобщей истории РАН
Тел: (095) 938-58-69

Среди требований, предъявляемых к разрабатываемой базе данных «Византийское право» [1], выделим следующие:

1. Возможность работы с многоязычными текстами, т.е. с текстами, написанными на нескольких языках. В частности необходима поддержка русского, латинского, греческого и старославянского языков.
2. Возможность осуществлять в тексте поиск слов и словосочетаний с учетом и без учета диакритических знаков. Это требование относится к греческому и старославянскому языкам.

Рассмотрим сначала техническую сторону построения кодировок, в равной степени относящуюся к шрифтам для любого языка и алфавита.

Начертания символов, физически присутствующих в шрифте, принято называть глифами (glyphs). Чтобы эти глифы были доступны в различных программах, каждому из них нужно приписать определенный код. Таким образом, под кодировкой шрифта понимается система соответствий между кодами символов и их начертаниями. До недавнего времени каждый символ обозначался одним байтом, что дает 256 возможных кодов. Однако для работы с многоязычными текстами этого явно недостаточно. Для того чтобы увеличить количество доступных кодов, необходимо увеличить количество байт, отводимых на один код. Вариантом такого решения является международный стандарт Unicode, использующий двухбайтовые коды [3].

Перейдем теперь к особенностям поиска в текстах, написанных на языках, использующих большое количество диакритических знаков. Для таких языков возможны два принципа составления кодировки [2]: на основе комбинируемых диакритических знаков (combining diacritics) и на основе предопределенных комбинаций (precomposed characters).

* Исследование выполнено при финансовой поддержке РГНФ. Проект № 02-05-12023 в.

В кодировках на основе комбинируемых диакритических знаков каждый диакритический знак оформляется как отдельный символ, смещенный за левый или правый край отведенного для него пространства. Самому этому пространству приписывается нулевая или ничтожно малая толщина. При печати такой символ перекрывается с предыдущим или последующим символом, что позволяет разместить акцент над или под буквой. Принципиальные достоинства этого метода:

- простота изготовления шрифта, т.к. количество символов, подлежащих включению в кодовую таблицу, сводится до минимума;
- простота ввода текста, т.к. небольшое количество символов можно разместить так, чтобы все они были доступны с английской клавиатуры;
- при условии соблюдения предыдущего принципа шрифт приобретает большую устойчивость к преобразованиям текста между форматами разных программ и переносу с платформы на платформу, поскольку символы, входящие в стандартный американский набор, при всех перекодировках остаются на своих местах.

Однако там, где требуется высокое типографское качество текста, проявляются недостатки присущие всем шрифтам такого рода, т.к. трудно добиться того, чтобы диакритические знаки подходили к разным по ширине буквам. В некоторые шрифты для этой цели включено два или более набора диакритических знаков, различающихся положением над символом.

Другой принцип – это кодировки на основе предопределенных комбинаций. В этом случае каждое сочетание буквы с тем или иным диакритическим знаком оформляется как отдельный символ. Для того чтобы пользоваться таким шрифтом, нужно иметь специальную программу, которая могла бы обеспечить ввод акцентированных символов при помощи более или менее удобных сочетаний клавиш.

Примером такой кодировки является греческая часть международной кодовой таблицы Unicode [3]. Так, например, символ α в таблице Unicode имеет код 03B1, а комбинированный символ ά - код 03AC.

Главным достоинством шрифтов на основе предопределенных комбинаций является их высокое типографское качество. Но они имеют значительно большее количество символов, чем кодировки на основе комбинируемых диакритических знаков, что создает дополнительные трудности при разработке раскладки клавиатуры под выбранный шрифт.

Основная трудность поиска в текстах на языках, использующих диакритические символы, заключается в том, что в этих языках слово может иметь несколько допустимых форм записи, различающихся только расстановкой диакритических знаков. Например, в дательном падеже множественного числа греческое слово πρᾶγμα (вещь) может иметь

формы πράγμασιν или πράγμασίν; слово πρόγονος (прародитель, предок) может иметь формы πρόγονος или προγονός. Если для представления такого текста использовать кодировку на основе предопределенных комбинаций, то каждая форма слова будет представлена своей, отличной от других, последовательностью кодов символов. Поэтому для того чтобы найти в тексте все допустимые формы искомого слова, необходимо последовательно выполнить поиск для каждой формы. Одним из способов решения этой задачи является исключение диакритических знаков из искомого слова и рассматриваемого текста в процессе поиска. Для осуществления этого преобразования надо в процедуре поиска для всех символов текста выполнять перевод комбинированных символов с диакритическими знаками в соответствующие символы без диакритических знаков. Такая процедура требует значительных временных затрат. При использовании кодировки, основанной на комбинируемых диакритических знаках, процесс исключения из текста диакритических знаков в процедуре поиска сводится к пропуску соответствующих символов.

Таким образом, при разработке базы данных «Византийское право», в соответствии с описанными требованиями, построена оригинальная кодировка, обладающая следующими основными характеристиками:

1. Каждый символ представляется двухбайтовым кодом.
2. Кодировка основана на принципе комбинируемых диакритических знаков.

Назначение байтов в коде символа: первый байт – код языка, второй байт – код символа в соответствующем языке.

В рамках базы данных «Византийское право» используются следующие языки:

- Русский. Используются коды символов и глифы шрифта Arial Cyr.
- Латинский. Используются коды символов и глифы шрифта Arial Cyr.
- Греческий. Используются коды символов и глифы шрифта Hellenica.
- Старославянский. Используются коды символов и глифы шрифта Cyrillica Bulgarian Phonetic.

В случае необходимости такая кодировка позволяет расширить номенклатуру используемых языков.

Для ввода и отображения многоязычных текстов были разработаны специальные элементы управления, пример использования которых показан на рисунке 1.

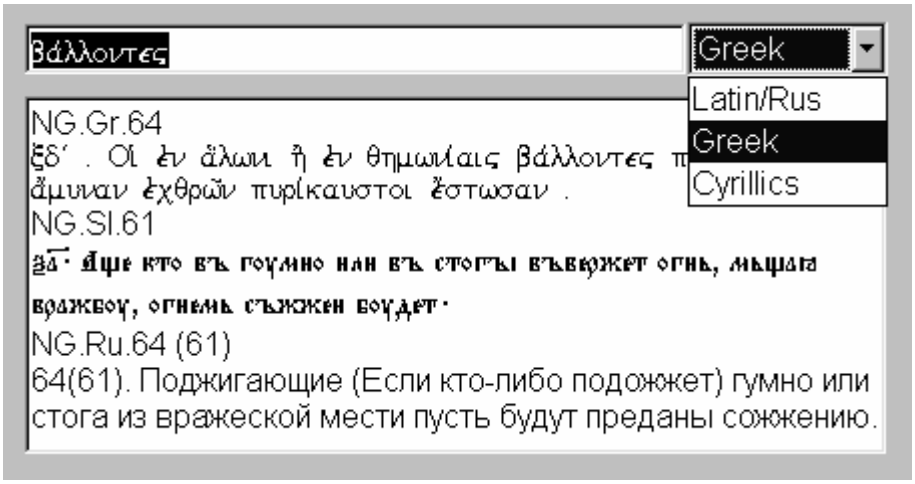


Рис.1. Элементы управления для ввода и отображения многоязычных текстов

Базовым для всех разработанных элементов управления является класс OurTextBox. Это контейнерный класс, содержащий объект ActiveX Microsoft Rich Textbox Control. Элемент Rich Textbox является окном отображения и редактирования текста в формате RTF. Этот формат дает возможность оперировать с многоязычными текстами, используя разные шрифты. Rich Textbox позволяет вводить, изменять, удалять текст; выделять и копировать часть текста; сохранять текст в файл и загружать его из файла. Для перевода текста из формата RTF в формат разработанной двухбайтовой кодировки класс OurTextBox содержит метод GetString. В этом методе перебираются все символы текста, содержащегося в поле Text элемента Rich Textbox, и к каждому символу добавляется байт кода языка, значение которого определяется значением поля SelFontName элемента Rich Textbox. Метод PutString осуществляет обратное преобразование.

В целом, разработанные кодировка и элементы управления отвечают предъявленным к ним требованиям и могут найти применение в различных системах, оперирующих с многоязычными текстами.

### *Список литературы*

1. Вин. Ю.Я., Гриднева А.Ю. База данных «Византийское право»: Итоги и перспективы // Круг идей: электронные ресурсы исторической информатики: Труды VIII конференции Ассоциации «История и компьютер» / Под ред. Л.И. Бородкина, В.Н. Владимирова. – М.; Барнаул: изд-во Алт. ун-та, 2003. – с.134-157.
2. http://www.basileia.narod.ru
3. http://www.unicode.org/charts/