Искомый компромисс, или О значении «таблиц Белградской группы»

А. В. Коваленин
Институт систем информатики СО РАН, Новосибирск

The tables of graphemes proposed by the Belgrad Group were part of the Proposal on the standardization of text encoding invoked by the vital necessity but that was not irreproachable conceptually. At the same time the tables have an independent scientific and applied value and represent, in particular, the form of the basis of another solution of the encoding task that does not contradict the Unicode concept.

Как уже не раз отмечалось, вопрос о создании единой кодировки для текстов на кириллической основе отличается по сложности от кодирования графической системы отдельной локально-временной письменной традиции, когда набор графем ограничен, а критерии их использования изучены. Это означает, что некоторые подходы, привычные для частных кодировок, перестают работать.

1. *Понимание стандарта*. Кодировки основных современных письменностей изначально разрабатывались исходя из метафоры пишущей машинки — то, что на машинке нажималось одной клавишей (после которой сдвигалась каретка), получало отдельный код. С накоплением опыта, однако, такой подход пришлось преодолеть, что было закреплено в требованиях Уникода. Кодируемая единица теперь не только *может* занимать несколько кодовых позиций, но это стало обязательным для составных символов. Даже позиционные варианты (как σ и ς) не должны кодироваться различно, раз их выбор определим контекстом. Видимые исключения связаны с «обратной совместимостью», то есть несут груз устаревших решений.

Важно помнить, что вопрос о кодировке важен главным образом для хранения текста, который придётся потом обрабатывать компьютеру, включая процесс отображения хранимого. Система отображения, как минимум, включает в себя шрифт,

но сегодня уже и сами шрифты поддерживают довольно сложную обработку цепочек кодов, что и делает Стандарт широко востребованным практически. Самим Стандартом вопрос о существовании требуемой системы не рассматривается; его цель — только указать устройство plain-text, то есть такой записи, которая однозначно и полно представляет текст, освобождённый от несодержательных графических нюансов. В этой освобождённости — смысл и польза кодировки. Примерами такой записи могут служить условная запись, применявшаяся в научных публикациях при отсутствии адекватного шрифта, или запись букв комбинациями более простых знаков. Например, plain-text для слова оу҃мъ может выглядеть *o_y=′mъ* (один подход к кодировке, первая буква занимает пять кодовых позиций), или *y3mъ* (другой подход; две позиции); лигатурное сочетание графем а и ѹ может записываться как *a&y* и т. п. Подобное представление уже сейчас стандартно, например, для символов с ударениями, которые кодируются последовательностью кода буквы и кода ударения.

2. *Основания кодировки*. Чтобы графема получила особый код, недостаточно её особого внешнего вида — надо ещё, чтобы такая особенность не была обусловлена почерком и не была позиционно определимой. Наконец, надо выявить её особую «функцию» в конкретной графической системе, которая отличает значащий графический вариант от незначащего. Однако основать всеобъемлющую практическую кодировку на таких «функциях» невозможно, потому что они могут быть не единственны в одной традиции и различаться в разных. Так, функцией буквы є в поморских рукописях является йотирование (в древних эту функцию несла ѥ), что позиционно определимо, то есть не требует отдельного кода; а в текстах позднего извода, кроме позиционного использования в начале слова, є получила ещё и семантическую нагрузку (указание мн. ч.). Единственное, что роднит использование одной графемы в разных традициях — это относимость к какой-то из основных букв кириллической азбуки (в данном случае «есть»), вариантом которой её можно считать в самом общем смысле.

3. *Противоречие*. Когда исследователь встречает новую графему, он ещё не знает, есть ли у неё та «функция», которая

делает её значимой для фиксации. Чтобы исследовать этот вопрос, надо ввести графему в корпус. Таким образом, с точки зрения науки, такие символы необходимо фиксировать; и даже при отрицательном решении вопроса учёт необычной графемы может быть полезен для изучения эволюции графических систем, так как наличие и частотность таких графем может служить дифференцирующим признаком.

Но с точки зрения Стандарта такие символы не должны фиксироваться, пока этот научный вопрос не решён положительно и доказательно. Противоречие обостряется тем, что такая доказательность может оказаться значимой для небольшого числа рукописей, вплоть до разового решения писца одной древней книги, а включение знака в Стандарт «давит» на всю систему.

Это противоречие требует принятия решений, не связанных с попытками уговорить консорциум выделить в стандарте кодовые позиции для очередного редкого символа. В связи с этим предлагалось шрифты для исследования рукописей подчинить внутреннему стандарту сообщества славистов, задействовав для этого кодовые позиции PUA (Private use area).

4. *Таблица Белградской группы*. Так или иначе, вопрос упирался в инвентаризацию графем. Такая работа была проделана группой, сложившейся после конференции в Белграде в октябре 2007 года (З. Костич, Х. Миклас, В. Баранов, В. Савич и др.), и представлена как Белградские предложения[1]. В ней, в частности, к каждой основной букве кириллицы были выписаны 1) «функциональные» варианты — для которых выявлены функции-основания для присвоения отдельного кода; 2) прочие графические варианты. Отдельным разделом таблицы представлялись лигатуры, и даже числовые выражения и прочие знаки.

В Белградских предложениях основные (включая уже стандартные) и функциональные буквы предлагалось поместить в PUA, «все выносные буквы зарегистрировать отдельно с титлом и без, отдельно строчные и прописные, для записи со сдви-

[1] Стандард старословенског ћириличког писма / коначни предлог, аутори Хајнц Миклас, Виктор А. Баранов, Зоран Костић, Виктор Савић. — Света Гора Атонска: Манастир Хиландар = Monastery Hilandar, 2008 (Београд: ИЦА). 24 стр. ISBN 978-86-84747-30-5.

гом влево и для записи между буквами. То же касается диакритических помет и титл». Так же отдельно предлагалось кодировать каждую лигатуру. Это предложение противоречит уже достигнутому в Уникоде пониманию, описанному выше. Оно ориентировано на упрощение средств визуализации хранимого, а не на удобство его последующего анализа.

5. *Предложение*. Предлагается другой подход. Его главное отличие — каждый нестандартный символ, даже монолитный на вид, занимает не одну, а две-три кодовые позиции. Первая из них — код «основной» буквы кириллической азбуки, остальные (функциональные и нет) содержат номер её варианта во внутреннем стандарте, принятом в научном сообществе (основной букве номер не нужен). В качестве формы предлагается взять Белградские предложения, которые являются профессиональным результатом работы филологов, отражающим представление о составе графем в славянской традиции. Если при этом все новые «функциональные» варианты букв поставить в начале ряда прочих графических вариантов, то все имеющиеся в таблице графемы можно использовать, не занимая позиций у Уникода. Это происходит следующим образом.

Если стандартная буква *b* имеет 20 графических вариантов, то они в plain-text имеют вид: *b1*, *b2*,…, *b20*. Цифры индексов здесь условно изображают символы-модификаторы, каждый из которых занимает свою кодовую позицию. Пример показывает, что естественно обойтись десятью модификаторами, но решение не обязательно основывается на десятичной системе. Конкретный выбор кодов для этих модификаторов может быть разным:

1) а) занять 10 позиций в PUA или б) использовать для модификаторов имеющиеся символы, которые так и выглядят индексами: (00BA (^{0}), 00B9 (^{1}), 00B2 (^{2}), 00B3 (^{3}), 00AA (^{a}), и другие символы, не используемые в славянском диапазоне, например, штрихи 2032-2037).

2) использовать 256 позиций, специально предусмотренных Уникодом для подобных модификаторов, из которых первые 16 находятся в основном кодовом пространстве (FE00–FE0F), остальные 240 — в Plane 14, или Supplementary Special-purpose Plane (E0100–E01EF). В таком случае модификатор всегда будет требовать одну кодовую позицию.

Отображение закодированного таким образом текста может быть технически поддержано (в частности, в устройстве шрифта) разными способами, в зависимости от целей представления, например:

1) в стандартном шрифте (типа Times):

а) игнорируя модификатор — то есть «a1 с ударением» должно отображаться: á. Так может отображаться зафиксированный текст для читателя, которому не важны графические нюансы.

б) показывая модификатор как индекс — то есть «a1 с ударением» должно отображаться: $á_1$. Так может отображаться тот же текст, например, для научной публикации, для которой важно отразить наличие графических различий и номера вариантов, но не их конкретный вид (например, при отсутствии специализированного шрифта или при сопоставительной выдаче текстов с разной графикой, что требует унификации шрифта);

2) в специализированном шрифте, отражающем графику представляемой традиции, содержатся требуемые начертания и инструкции по замене на них комбинаций буквы с модификатором; комбинации, отсутствующие в отображаемой шрифтом традиции, могут отображаться по какому-то из вариантов п.1.

6. *Следствия*. Такая форма хранения текста позволяет максимально точно учитывать графические нюансы и в то же время легко от них отвлекаться путём специального шрифта или программного удаления неважных для текущей задачи модификаторов. И наоборот, предложенное решение позволяет организовать поэтапное кодирование, при котором уточнение графем производится добавлением модификаторов.

Использование диапазона модификаторов позволяет не использовать PUA, полностью вписываясь в стандарт. Поскольку нам ещё неизвестны системы кодирования славянских текстов, использующих модификаторы, можно строить кодировочное решение, свободное от непоследовательности прошлых решений. Можно даже предусматривать дополнительное отображение через модификаторы некоторых стандартных символов, которые удобно считать вариантом основной буквы (например, трактовать ϵ как вариант «есть»), чтобы можно было впоследствии плавно перейти на более «концептуально чистое» решение.

Решение делает открытым перечень вариантов букв, которые возможно фиксировать, позволяет легко фиксировать окказиональные варианты до их утверждения. Само утверждение перечня становится прерогативой научного сообщества (например, Конгресса славистов). А Белградские предложения, являясь образцом требуемой для этой задачи формы решения, должны быть ещё раз внимательно рассмотрены в свете предложенной техники кодирования.