ИСПОЛЬЗОВАНИЕ НЕСТАНДАРТНОЙ КОДИРОВКИ В ИНТЕРНЕТЕ: ВЗГЛЯД НА ПРОБЛЕМУ

А. В. ЧЕМЫШЕВ,
инженер-программист Центра инновационных языковых технологий ГАОУ ВПО «Коми республиканская академия государственной службы и управления»
(г. Сыктывкар, РФ)

Статья посвящена проблеме использования национальными газетами, журналами, книгоиздателями нестандартных шрифтов с различными «устаревшими» кодировками и последствиям, к которым их использование приводит.

• *информационная цивилизация; шрифт; кодировка; национальный алфавит; Юникод*

Еще со времен И. Гутенберга, когда появилась возможность печатать книги, тиражировать новости, шрифты (тогда наборные шрифты) были важным моментом обеспечения письменности любого народа. Этот перелом информационных цивилизаций разделил народы на письменные и бесписьменные. У бесписьменных народов язык стал функционировать только на бытовом уровне, постепенно вымывался и исчезал.

До следующего перелома информационных цивилизаций языкам письменных народов мало что грозило. Сначала появились книги, газеты, журналы. Позже – радио, телевидение и другие средства массовой информации. Появление компьютеров значительно облегчило работу по набору, редактированию, обработке текстов. После принятия компьютеров на «вооружение» полиграфистами, газетчиками, журналистами для них нужно было изготовить всевозможное количество разных компьютерных шрифтов. А у разных народов разное графическое представление национальных алфавитов. Для шрифтов предусмотрены различные способы кодирования, или кодировки. Кодировка зависит от того, в какой системе и для каких задач будут применяться данные шрифты. Если шрифты предназначены для использования в локальной работе, например, чтобы набрать текст, сверстать газету и вывести ее на пленку для дальнейшего тиражирования, то о кодировке шрифта можно не задумываться. В 3, 4 и даже 10 компьютерах, задействованных в данной технологической цепочке, можно установить именно этот шрифт, даже не зная о том, какая кодировка там используется.

С появлением угрозы ядерной войны после Карибского кризиса 1960-х гг. по заказу американских военных началась разработка катастрофоустойчивых вычислительных комплексов путем создания территориально-распределенной сети. Угроза ядерной войны миновала, и к этой сети начали подключаться простые американцы. Так появился Интернет. Вместе с ним настала новая эпоха перелома информационных цивилизаций.

В первую очередь в Интернет «попросились» Великобритания, Австралия и Новая Зеландия. С подключением Европы он перестал быть только англоязычным. За Европой последовали остальные континенты, в начале 1990-х гг. – Советский Союз, а потом и Россия. Весь земной шар стал окутан Всемирной паутиной, и

Интернет превратился в культурный феномен, буквально на глазах изменивший и продолжающий изменять все без исключения стороны человеческой жизни.

В мире более 6 000 языков, для графического представления национальных алфавитов используются различные знаки на основе латиницы, кириллицы, различные иероглифы и т. д. До начала 1990-х гг. емкость большинства используемых кодировок (ASCII, КОИ-8, Microsoft CP1251 и т. д.) не превышала 256. Например, для большинства языков на территории Советского Союза, а позже России использовалась кодировка КОИ-8. Для всех знаков национальных алфавитов с учетом прописных и строчных букв места в кодовой таблице не хватало. Выход был найден простой: в изготавливаемых шрифтах знаки национального алфавита ҷҶ, ӧӦ, йЙ, ӟӞ, жЖ, ҥҤ, ўЎ, ӓӒ, ыЫ и т. д. подменялись знаками ҷҶ, ќЌ, ћЋ, њЊ, љЉ, ‰ˆ, єЄ, ўЎ, іІ, ї Ї, ¢™ и др. В выборе знака для новой буквы не было никакой системы и логики. Более того, в одном и том же языке в 5 разных шрифтах один и тот же знак мог быть закодирован 5 способами. Например, в марийском языке буква ӱӰ могла быть передана знаком ¢™, ўЎ или каким-либо другим. Такое положение дел было обусловлено тем, что шрифты изготавливались полукустарным способом не имеющими никакого специального образования и подготовки людьми, часто на ворованном программном обеспечении и с нарушением авторских прав правообладателей шрифтов, на основе которых национальные шрифты делались. На сегодняшний день, вдобавок, эти шрифты не удовлетворяют требованиям СанПиН: они портят зрение детей – с их помощью запрещено выпускать учебники и детскую литературу. Однако по всей России с использованием этих кустарных шрифтов до сих пор выпускаются марийские газеты, удмуртские журналы, коми книги и т. д.

Как же быть, если с помощью таких шрифтов вести сайты на марийском языке, блоги на удмуртском, обмениваться электронной почтой, мгновенными сообщениями на коми и других языках? Ответ прост: никак. Вместо национальных букв там будут непонятные символы.

В 1991 г. некоммерческой организацией «Консорциум Юникода» был предложен единый стандарт для представления знаков алфавитов всех языков мира. Этот стандарт получил название «Юникод» или «Уникод» (UTF-8, UTF-16 и UTF-32 – разновидности Юникода; в данный момент широкое распространение получил UTF-8, хотя компания Microsoft также использует UTF-16). Основной принцип Юникода – каждому уникальному знаку принадлежит единственный код, например, Ӓ, ӓ – 04D2, 04D3; Ҥ, ҥ – 04A4, 04A5; Ӧ, ӧ – 04E6, 04E7; Ӯ, ӯ – 04F0, 04F1; Ӹ, ӹ – 04F8, 04F9; Ӝ, ӝ – 04DC, 04DD; Ӟ, ӟ – 04DE, 04DF, Й, й – 04E4, 04E5; Ҷ, ҷ – 04F4, 04F5, I, i (в кириллической зоне) – 0406, 0456 (I, i латиницы имеют другой код) и т. д. Правда, присвоение кодов буквам и знакам различных алфавитов не обошлось без курьезов: например, «Консорциум Юникода» не присвоил коды около 38 буквам, используемым в письменности народов России. Эти 38 букв «не видны» – не отображаются в Интернете. Мириться с таким положением дел нельзя, необходимо добиваться «признания» «Консорциумом Юникода» этих знаков и присвоения им кода. Вопрос в том, кто должен заниматься подготовкой и подачей заявки на включение «забытых» букв в Юникод? Для обычных граждан или организаций этот процесс может длиться несколько лет, если же к ним (к «Консорциуму») обращается государственный орган (например, Министерство связи и массовых коммуникаций Российской Федерации), то рассмотрение подобных вопросов происходит очень быстро и оперативно. Те народы, чьих букв не оказалось в Юникоде, с присвоением кода могут начать пользоваться Интернетом на родном языке. Пока им дорога туда, к сожалению, закрыта.

После появления Юникода долгое время не было шрифтов, поддерживающих данный стандарт, или то количество созданных шрифтов не устраивало

национальных полиграфистов и средства массовой информации. Газеты, журналы и книги продолжали выпускаться с использованием нестандартных шрифтов. Производители программного обеспечения не спешили с переходом на Юникод – компания Microsoft до появления своего Windows 7 поддерживала его только частично и чаще декларативно. Этого не скажешь о Unix-подобных системах – в Linux всегда поддерживался UTF-8.

В данный момент почти все крупные производители шрифтов выпускают юникод-шрифты. Много лицензионных шрифтов под открытой лицензией: Charis SIL, Doulos SIL, PT Sans, PT Serif, PT Mono. Последние три шрифта выпущены российским лидером в области изготовления коммерческих шрифтов – фирмой «ПараТайп». Бесплатные шрифты от «ПараТайп» используются и рекомендованы мировыми IT-гигантами Google и Apple Inc. Примечательно, что сайт премьер-министра Соединенного Королевства Великобритании и Северной Ирландии, главы правительства, главы государства, главного советника монарха Дэвида Кэмерона оформлен бесплатными шрифтами фирмы «ПараТайп» PT Sans и PT Serif, выпущенными при финансовом участии Роспечати РФ для бесплатного распространения среди народов и народностей России.

Для обычных пользователей количество бесплатных лицензионных юникод-шрифтов достаточное. Но это количество, конечно же, не устраивает национальных издателей. Однако эта проблема решаема: у того же «ПараТайп» можно приобрести лицензии на использование коммерческих шрифтов. У них есть пакеты по 50, 100, 300 и более шрифтов. Регионы могут централизованно закупить их для своих национальных издательств. В XXI в. всем необходимо понять, что контрафактные шрифты вносят неразбериху некорректным отображением букв национального алфавита в Интернете и к тому же портят зрение детей!

Mari-Arial, Mari-Baltica, Mari-Courier, Mari-Futura, Mari-Helvetica, Mari-Journal, Mari-New York, Mari-Palette, Mari-Park Avenue, Mari-Parsek, Mari-Peterburg, Mari-Plain, Mari-Plakat, Mari-Poster, Mari-Pragmatica, Mari-SchoolBook, Mari-Taurus, Mari-Time Roman, Mari-TimesET и т. д. – всего подобных шрифтов около 300. Несколько меньше шрифтов со словами «Komi»: Komi Academy, Komi Antiqua, Komi Arbat, Komi Arial, Komi SchoolBook, Komi TextBook – и «Udm»: Arial Udm, Courier New Udm, Times New Roman Udm, Verdana Udm.

В 1991 г. некоммерческой организацией «Консорциум Юникода» был предложен единый стандарт для представления знаков алфавитов всех языков мира. Этот стандарт получил название «Юникод» или «Уникод». Основной принцип Юникода – каждому уникальному знаку принадлежит единственный код.

Шрифты Mari-... и Komi… изготавливались «народными умельцами» в середине 90-х гг. с помощью незатейливых редакторов на основе существующих шрифтов, как правило, фирмы «MonoType» (конечно же, не без нарушения авторских прав правообладателей шрифтов) и делались от безысходности: национальные газеты, журналы, книги необходимо было выпускать, набор текстов уже осуществлялся на компьютерах (в основном тогда использовался Windows 95, позже Windows 98 и XP), а соответствующих компьютерных шрифтов не было. Такие «шалости» можно списать на «лихие» 90-е. Парадоксально, что нестандартные контрафактные шрифты Arial Udm, Courier New Udm, Times New Roman Udm, Verdana Udm были заказаны в 2006 (!) г., когда уже всем было понятно, что национальные шрифты должны быть «юникод-совместимыми», и были заказаны... по поручению властей Удмуртской Республики за бюджетные деньги (!!!) фирме «Градиент – Новые технологии». Бюджет по разработке, созданию и цен-

трализованному распространению шрифтов Arial Udm, Courier New Udm, Times New Roman Udm, Verdana Udm в Удмуртской Республике на компакт-дисках, а также на сайтах государственных учреждений (Государственный Совет УР, Министерство национальной политики УР и др.) составил 2 млн руб.! Разработка фирмой «Градиент – Новые технологии» и дальнейшее распространение набора суррогатных удмуртских шрифтов производились нелегально, поскольку лицензионное соглашение конечного пользователя (EULA), по которому корпорацией «MonoType» предоставлялись файлы шрифтов Arial, Courier New, Times New Roman и Verdana, использованных фирмой «Градиент – Новые технологии», не предусматривало возможности внесения пользователем каких-либо изменений в шрифтовые продукты, а сами названия Arial, Courier New, Times New Roman и Verdana являются зарегистрированными торговыми марками корпорации «MonoType».

Использование не соответствующих международному стандарту Юникод национальных шрифтов, поддерживающих произвольные кодировки, идет вразрез с общемировыми и российскими тенденциями в области информатизации, уже сейчас отбрасывая пользователя в области технических решений на 10–15 лет назад. В дальнейшем этот разрыв будет лишь расти.

В шрифтах Arial Udm, Courier New Udm, Times New Roman Udm, Verdana Udm удмуртские буквы находятся на месте некоторых южнославянских букв и в Интернете отображаются как Љљ, Ќќ, Џџ, Ћћ и Њњ [1; 2]. С помощью таких шрифтов оформлены многие государственные сайты Удмуртской Республики, например, сайт Государственного Совета УР: http://www.udmgossovet.ru. Попытаемся на указанном сайте прочитать новость на удмуртском языке. Вот что мы увидим: «Со пќртэм улосъесын но кунъесын улћсь, пќртэм кылъесын вераськись но одћг выжыосты утись калыкъесты огазеяны, соослэсь аспќртэмлыксэс возьыны но туала вакытэ пќртэм ќръесъя азинскыны юрттыны кулэ...» А должно быть: «Со пӧртэм улосъесын но кунъесын улӥсь, пӧртэм кылъесын вераськись но одӥг выжыосты утись калыкъесты огазеяны, соослэсь аспӧртэмлыксэс возьыны но туала вакытэ пӧртэм ӧръесъя азинскыны юрттыны кулэ...»

Другой пример: учреждение, находящееся в ведении Министерства культуры, печати и по делам национальностей Республики Марий Эл, – Республиканский научно-методический центр народного творчества и культурно-досуговой деятельности пишет в филиал Государственного Российского Дома народного творчества «Финно-угорский культурный центр Российской Федерации» электронное письмо с марийским текстом для размещения новости на интернет-портале или для титров в создаваемых Центром документальных фильмах (например, http://fusee.tv/heading/10/). В письме встречаются такие слова: «Эре‰ер», «Т¢рлем¢дыр», «Унчо ¢жара», «Шанавїл». Как человек без знания марийского языка должен догадаться, что при написании вышеуказанных слов автор имел в виду «Эреҥер», «Тӱрлемӱдыр», «Унчо ӱжара» и «ШанавӚл»? Не проще ли было автору написать в международной стандартной кодировке Юникод? Все бы правильно отображалось!

Использование не соответствующих международному стандарту Юникод национальных шрифтов, поддерживающих произвольные кодировки, идет вразрез с общемировыми и российскими тенденциями в области информатизации, уже сейчас отбрасывая пользователя в области технических решений на 10–15 лет назад. В дальнейшем этот разрыв будет лишь расти. Невозможно рационально обосновать практику использования устаревших, не соответствующих стандарту суррогатных решений при наличии

доступных, отвечающих современным техническим требованиям профессиональных решений.

За последние несколько лет были созданы достаточно качественные, типографски полноценные шрифты под различными видами свободных лицензий с поддержкой всех языков мира, соответствующие стандарту Юникод [3]. В дальнейшем их число, несомненно, будет увеличиваться. Однако эти шрифты созданы зарубежными разработчиками и, как следствие, не учитывают традиции российской типографии. Из-за особенностей дизайна они не могут считаться универсально применимыми. Кроме того, задача поддержки языков народов Российской Федерации создателями этих шрифтов специально не ставилась.

С целью преодоления перечисленных выше недостатков в 2009 и 2010 гг. российской фирмой «ПараТайп» были созданы свободные шрифты PT (PT Sans, PT Serif, а в 2012 г. еще PT Mono), поддерживающие все языки России в соответствии со стандартом Юникод. PT Sans, PT Serif и PT Mono являются шрифтами универсального назначения с открытой пользовательской лицензией. Они призваны не только обслуживать печатные издания, сетевые информационные ресурсы, официальную и деловую переписку, образование и науку, но и способствовать развитию национальных письменностей и межкультурного обмена. Фирмой «ПараТайп» проведена масштабная работа по изучению потребностей языков народов Российской Федерации в шрифтовой поддержке, исследованию сложившихся в традиционной типографике на языках этих народов традиций.

Маниакальное упорство людей, продолжающих писать электронные тексты на удмуртском, марийском, коми и других языках с использованием нестандартной кодировки, поражает. Ведь эти тексты попадают сейчас в Интернет.

Что же надо сделать, чтобы избавиться от наследия 90-х гг.?

Первое. На мой взгляд, во всех регионах Российской Федерации необходимо принять или постановления местных правительств, или решения терминоорфографических комиссий об обязательном использовании стандарта Юникод в качестве единого стандарта для кодирования национальных букв. Такое постановление готовится в Республике Коми. В Якутии оно принято несколько лет назад.

Второе, и очень важное. Необходимо признать неправильное отображение национальных символов в словах, фразах, предложениях в Интернете, компьютерной технике за грамматические ошибки автора текста. Тогда того, кто пишет в Интернете с использованием нестандартной кодировки, можно будет считать неграмотным человеком!

Поступила 21.04.2012

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Сахарных, Д. М.* Удмуртский язык при работе на персональном компьютере : методические рекомендации / Д. М. Сахарных, В. П. Аркашев. – Ижевск : ИПК и ПРО УР, 2010. – 12 с. – (Электронная версия публикации: http://files.udmurt.info/udmurtskiy-yazyk-na-kompjutere.pdf).
2. *Удмуртская* письменность. Вып. 1. Проблемы электронной письменности : науч.-метод. сб. [ред.: Д. М. Сахарных]. – Ижевск : ИПК и ПРО УР, 2011. – 12 с. – (Электронная версия публикации: http://files.udmurt.info/udmurtskaya-pismennost-vypusk-i.pdf).
3. *Языки* меньшинств в компьютерных технологиях: опыт, задачи и перспективы : сб. материалов междунар. конф. – Йошкар-Ола, 2011. – 96 с.