

Кодировка ASCII-Cyrillic и ее конвертор email-ru.tex

— Краткое введение —

Здесь приводится новое представление стандартной кодировки ASCII для русского алфавита, позволяющее набирать и читать тексты на русском языке при отсутствии русской клавиатуры или русских шрифтов – ситуация, часто встречающаяся за пределами России.

Предложенная кодировка ASCII-Cyrillic может использоваться как для русского, так и для украинского языка. Это краткое введение первоначально было написано для кодировки русского языка, однако впоследствии в кодировку были внесены изменения с целью ее адаптации к украинскому алфавиту.

Ниже приведен фрагмент русского сообщения по электронной почте. Сегодня такое сообщение в системе электронной почты можно представить в виде последовательности “октетов” или “байтов” (имеется в виду 8-символьный ряд из нулей и единиц), каждый из которых соответствует определенной букве в зависимости от принятой 8-битовой кодировки. В настоящее время большинство таких электронных сообщений, введенных на русской клавиатуре, затем можно прочесть (с помощью любого 8-битного русского экранного шрифта) на большинстве компьютеров в тех странах, алфавитом которых является кириллица – но оно вряд ли будет прочитано в других странах.

На обратном пути Michele объяснила мне, как делать пересадку на метро. Мы с ней проехали большую часть пути вместе. Она вышла на остановке после того, как мы пересели на мою линию. Пользоваться метро в Париже, действительно, очень просто -- гораздо проще, чем в Москве (чаще всего я пользовалась линией 13). Когда я это поняла, то сразу успокоилась. Сейчас всё в порядке. Я могу пользоваться метро, и уже не боюсь ходить по Парижу.

(Изображение в формате GIF, которое вы здесь видите на своем браузере, будет прочитано во всех странах, но оно, как минимум, в 10 раз более громоздкое и не очень четкое.)

Необходимость перекодировки 8-битового кириллического текста для принимающей операционной системы затрудняет его передачу по электронной почте. Когда кодировка принимающей стороны не содержит всех используемых букв, перекодировка текста может стать не только затруднительной, но и невозможной.

Утилита "email-ru.tex" конвертирует вышеприведенный 8-битный текст в новую 7-битную транскрипцию русского алфавита (ASCII-Cyrillic) и обратно. Назначение этой программы – ввод и чтение русских текстов на любых компьютерах, где бы эти компьютеры ни находились:

```
Na obratnom puti !Michele obq'asnila mne, kak
delath peresadku na metro. My s nej proexali
bolhwu'u 'casth puti vmeste. Ona vywla na
ostanovke posle togo, kak my pereseli na mo'u
lini'u. Polhzovaths'a metro v Pari'ze,
dejstviteljno, o'cenh prosto -- gorazdo pro'we,
'sem v Moskve ('ca'we vsego 'a polhzovalash
liniej 'N13!). Kogda 'a 'eto pon'ala, to srazu
uspokoilash. Sej'cas vs'o v por'adke. 'A mogu
polhzovaths'a metro, i u'ze ne bo'ush hodith po
Pari'zu.
```

На самом деле особенность этого письма, посланного по электронной почте в Россию, заключается в том, что оно было набранно в Париже на латинской (французской) клавиатуре в кодировке ASCII-Cyrillic, так как в Париже под рукой не было клавиатуры с кириллицей.

Если бы русская клавиатура была доступна, то можно было бы получить и послать обе версии письма (в кириллической кодировке и кодировке ASCII-Cyrillic). Далее, если бы версия в 8-битной кириллической кодировке была бы испорчена в процессе передачи или дешифровки, то версия в кодировке ASCII-Cyrillic могла бы быть прочитана и/или использована в России для получения удобной версии в 8-битной кириллической кодировке.

Обратите внимание на то, что в кодировке ASCII-Cyrillic для обозначения большинства русских букв взяты соответствующие буквы английского (латинского) алфавита, некоторые из которых (с предшествующим знаком ударения “’”) используются для обозначения оставшихся русских букв. В частности, русский кириллический алфавит из 33 букв:

```
а б в г д е ё ж з и й к л м н о п р
с т у ф х ц ч ш щ ъ ы ь э ю я
```

в кодировке ASCII-Cyrillic представлен следующим образом:

```
a b v g d e 'o 'z z i j k l m n o p r  
s t u f x 't c 's 'w q y h 'e 'u 'a
```

В дополнение к этому, вставка английского текста предваряется восклицательным знаком “!”. Перечисленные правила настолько просты, что, с большой вероятностью, ввод и чтение русского текста в кодировке ASCII-Cyrillic могут быть освоены в течение часа и доведены до совершенства в недельный срок.

Особое внимание надо обратить на тот существенный факт, что все символы, используемые в ASCII-Cyrillic, являются 7-битными (иными словами, 8-й бит соответствующего октета всегда будет нулевым), что позволяет использовать фиксированные значение и форму, задаваемые универсальным ASCII стандартом. Кроме того, все 8-битные кодировки кириллического текста не нарушают ASCII стандарт там, где речь идет о 7-битных символах.

В 7-битной кодировке ASCII-Cyrillic количество символов в русском тексте увеличится (но не более чем на 4 процента) по сравнению с аналогичным текстом в 8-битной кодировке. Таким образом, скорость ввода для кодировки ASCII-Cyrillic на любой клавиатуре может быть сравнима со скоростью ввода на кириллической клавиатуре.

При использовании современного “gzip”-сжатия обоих текстов разница их объемов, составляющая 4 процента, уменьшится до 1 процента. Следовательно, хранение кириллических текстовых файлов в кодировке ASCII-Cyrillic возможно без всяких отрицательных последствий.

Поскольку конвертор “email-ru.tex” позволяет выполнять обратное преобразование 7-битной кодировки ASCII-Cyrillic в любую 8-битную кодировку, имеется возможность двухшагового преобразования из одной 8-битной кодировки в другую.

Кодировка ASCII-Cyrillic является родственной существующим транскрипциям русского алфавита, которые расходятся в использовании концепции лигатуры (соединение из двух-трех английских букв, обозначающих определенную букву русского алфавита). Утилита “email-ru.tex” также конвертирует русский алфавит в общепризнанную систему лигатурных транскрипций, которая принята библиотекой Конгресса США:

```
Na obratnom puti Michele ob'jasnila mne, kak  
delat' peresadku na metro. My s nej proexali  
bol'shuju chast' puti vmeste. Ona vyshla na  
ostanovke posle togo, kak my pereseli na moju  
liniju. Pol'zovat'sja metro v Parizhe,  
dejstvitel'no, ochen' prosto -- gorazdo proshche,  
chem v Moskve (chashche vsego ja pol'zovalas'
```

liniej No13). Kogda ja eto ponjala, to srazu uspokoilas'. Sejchas vse v porjadke. Ja mogu pol'zovat'sja metro, i uzhe ne bojus' xodit' po Parizhu.

Замечание: Точное обратное преобразование в 8-битный формат существующих лигатурных транскрипций требует больших затрат времени.

Нельзя сказать, что представление ASCII-Cyrillic написано лучше, но его преимущество в том, что как для машин, так и для людей, оно абсолютно однозначно и удобно для чтения. Утилита “email-ru.tex” выполняет перевод текста в **обоих** направлениях без вмешательства человека. Преобразование (8-bit) \Rightarrow (7-bit) \Rightarrow (8-bit) **точно** возвращает оригинал 8-битного русского текста (с одной незначительной особенностью: во всех строках будут удалены заключительные пробелы).

Таким образом, используя кодировку ASCII-Cyrillic, русский текстовый файл можно удобно и надежно архивировать и передавать по электронной почте – отсюда и само название утилиты: “email-ru”.

Начальные рабочие инструкции по использованию “email-ru.tex” как конвертора довольно просты:

- Поместите копию конвертируемого файла рядом с “email-ru.tex” и задайте ему имя “in.txt”.
- Запустите email-ru.tex (не “in.txt”) с помощью Plain TeX. Командная строка в этом случае будет иметь вид: `tex email-ru.tex`
- Следуйте инструкциям “email-ru.tex”, появляющимся на экране монитора.

Полная техническая документация по ASCII-Cyrillic в настоящее время содержится **внутри** конвертора “email-ru.tex” для обеспечения его автономности. Следует сказать, что современный HTML-формат, возможно, более удобен для чтения благодаря тому, что в нем буквы кириллического алфавита выполнены с помощью общепризнанной GIF графики. (Смотрите также соответствующую PDF-версию.)

Примечание. Несколько важных TeX разработок, особенно “ctex” под unix и большинство разработок для операционной системы Macintosh OS, в настоящее время не позволяют выводить (командой TeX’a “\write”) правильные октеты со значения, большего 127 – как этого требует “email-ru.tex” при преобразовании ASCII-Cyrillic в 8-битный кириллический текст. (Эта проблема **не влияет** на преобразование 8-битного кириллического текста в ASCII-Cyrillic.)

Для решения этой проблемы пакет ASCII-Cyrillic снабжен небольшой автономной и переносимой утилитой “Kto8”, которая преобразует любой текстовый файл, получаемый на выходе некоторых «проблемных» Т_ЕX разработок, в правильный 8-битный текст.

Признаком того, что следует использовать эту утилиту, является наличие в выходном тексте конвертора “email-ru.tex” большого количества символов шляпки “^^”.

Для операционных систем Linux, Unix, Macintosh и Windows со временем могут появиться более совершенные бинарные версии “Kto8”. Здесь находится последняя версия утилиты “Kto8”. См. также СТАН-архив.

Более полная информация о кодировке ASCII-Cyrillic.

- **Директория программного обеспечения ASCII-Cyrillic** (создана в декабре 2000):

<http://topo.math.u-psud.fr/~lcs/ASCII-Cyrillic/>

- **Web страница ASCII-Cyrillic.** См. файл “ascii-cy.htm”.
- **Архивирование длинных терминов.** См. СТАН Т_ЕX архив и его зеркала.
- **Автор.** Laurent Siebenmann, CNRS, Франция; мэйлы посылайте по адресу: lcs@topo.math.u-psud.fr
- **Условия копирайт.** Gnu Public Licence.
- **Документация.** – в настоящее время включена в конвертер “email-ru.tex” в виде ASCII текста.
- **Русский перевод** (сентябрь 2001). Галина Горячевских, Марина Новожилова, Александр Брюханов.