ASCII-Cyrillic and its converter email-ru.tex

by Laurent Siebenmann

A new faithful ASCII representation for Russian called ASCII-Cyrillic is presented here, one which permits accurate typing and reading of Russian where no Russian keyboard or font is available -- as often occurs outside of Russia.

ASCII-Cyrillic serves the Russian and Ukrainian languages in parallel. This article initially discusses Russian; but, further along, come the modifications needed to adapt to the Ukrainian alphabet.

TeX is mother to ASCII-Cyrillic inasmuch as its converter "email-ru.tex" is a program in TeX language which is interpreted by TeX. On the other hand, ASCII-Cyrillic is designed to be useful beyond TeX. Indeed a current project is to make it available on Internet so that the vast public ignorant of TeX can exploit it. This provisional Internet bias has led to the use of HTML for this article, bitmaps and all.

Contents

Below is a photo of a fragment of Russian email sent from France to Russia. Ideally, it would be typed as 8-bit text using a Russian keyboard and screen font, and then assigned a suitable MIME type identifying the font encoding. To the email system, the message contents would be a sequence of "octets" or "bytes" (each 8 zeros or ones), where each octet corresponds to a character according to the font encoding. The

receiving email system and email reader are expected to recognize the encoding and provide for Cyrillic display and printing. This system works well provided there is diligent support of it from one end of the email trajectory to the other. The transcoding provided by "email-ru.tex" can be part of this support.

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 №13A,
действительно, очень просто -- гораздо проще,
чем в Москве. Когда я это поняла, то сразу
успокоилась. Сейчас всё в порядке. Я могу
пользоваться метро, и уже не боюсь ходить в
Париже.

(The GIF photo image you see here is widely readable, but at least 10 times as bulky as 8-bit text, and somewhat hazy too.)

Unfortunately, quite a few things can go wrong in MIME-tagged 8-bit Cyrillic email, particularly when either sender or recipient is outside the countries using a Cyrillic alphabet:

-- there is a frequent need to re-encode for another computer operating system, and when the targeted encoding does not contain all the characters used, defects result. Worse, if at any stage wrong assumptions are made about an encoding, severe and irreparable damage usually ensues.

-- outside the Cyrillic world, Cyrillic keyboards are rarissime, and Cyrillic screen fonts often have to be hunted down and installed by an expert.

To circumvent such difficulties Russian speakers often use an ad hoc 7-bit ASCII transliteration of Russian (or even switch to English) and then rely on ASCII's universal portability. ASCII, the American Standard for Computer Information Interchange of 1963, long predates Internet and the MIME protocols.

ASCII-Cyrillic is a new faithful ASCII transcription of Russian that transforms this "last resort" ASCII approach into something quite general and reliable.

For example, the email fragment illustrated above was first typed as the ASCII-Cyrillic text below, using a Latin (French) keyboard, and then converted to the above Cyrillic form by the utility "email-ru.tex". For good measure, both forms were then emailed.

```
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 'N13!A,
dejstvitelhno, o'cenh prosto -- gorazdo pro'we,
'cem v Moskve. 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 xodith v
Pari'ze.
```

Inversely, for email from Russia to France, the keyboarding would be Cyrillic and "email-ru.tex" would convert from 8-bit text to ASCII-Cyrillic. Again, for good measure, both versions would be sent.

ASCII-Cyrillic is designed to be both typeable and readable on every computer worldwide: Well chosen ASCII letters stand for most Russian letters. To distinguish the remaining handful of Russian letters, a prefixed accent ' is used. Further, to introduce English words, the exclamation mark ! appears. The rules are so simple that, hopefully, ASCII-Cyrillic typing and reading of Russian can be learned in an hour, and perfected in a week.

An essential technical fact to retain is that all the characters used by ASCII-Cyrillic are 7-bit (i.e. the 8th bit of the corresponding octet is zero), and each character has a reasonably well-defined meaning and shape governed by the universally used ASCII standard. It is a key fact that all 8-bit Cyrillic text encodings include and respect the ASCII standard where 7-bit characters are concerned.

In 7-bit ASCII-Cyrillic form, Russian prose is about 5 percent bulkier than when 8-bit encoded. Thus, typing speed for ASCII-Cyrillic on any computer keyboard can approach that for a Cyrillic keyboard.

The difference of 5 percent in bulk drops to about 1 or 2 percent when the common "gzip" compression is applied to both. Thus, there is virtually no penalty for storing Cyrillic text files in ASCII-Cyrillic form.

As "email-ru.tex" converts both to and from ASCII-Cyrillic, one can convert in two steps between any two common 8-bit Cyrillic encodings. Further, new or "variant", 8-bit encodings can be quickly introduced "on-the-fly" by specifying an "encoding vector". Additionally, the Cyrillic UTF8 unicode will soon be supported.

ASCII-Cyrillic is a cousin of existing transcriptions of Russian which differ in using the concept of ligature -- i.e. they use two or more English letters for certain Russian letters. The utility "email-ru.tex" also converts Russian to one such ligature-based transcription system established by the the USA Library of Congress:

```
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 No13A,
dejstvitel'no, ochen' prosto -- gorazdo proshche,
chem  v Moskve. Kogda ja eto ponjala, to srazu
uspokoilas'. Sejchas vse v porjadke. Ja mogu
pol'zovat'sja metro, i uzhe ne bojus' xodit' v
Parizhe.
```

Nota bene:- Accurate reconversion of existing ligature-based transcriptions back to 8-bit format always requires a good deal of human intervention.

Although not more readable, the ASCII-Cyrillic representation has the advantage that, for machines as well as men, it is completely unambiguous as well as easily readable. The "email-ru.tex" utility does the translation both ways without human intervention, and the conversion (8-bit) ==> (7-bit) ==> (8-bit) gives back exactly the original 8-bit Russian text. (One minor oddity to remember: terminal spaces on all lines are ignored.)

Thus, by ASCII-Cyrillic encoding a Russian text file, one can archive and transfer it conveniently and safely, even by email.

Beginner's operating instructions

To use "email-ru.tex" as a converter:

- Put a copy of the file to convert, alongside of "email-ru.tex" and give it the name "IN.txt".

- Process "email-ru.tex" (not "IN.txt") with Plain TeX. The usual command line is: `tex email-ru.tex`

- Follow the instructions then offered on screen by "email-ru.tex".

A batch mode will soon be available.

Use of ASCII-Cyrillic with TeX

ASCII-Cyrillic could be made completely independent of TeX through using, to build its converter, some other portable language (C, Java, ...). On the other hand, the TeX community, with its keen appreciation of ASCII text as a stable portable medium, will probably always be "home ground" for ASCII-Cyrillic. Thus, it is unfortunate that, for lack of time, this author has not so far created macro packages offering optimal integration of ASCII-Cyrillic into Cyrillic (La)TeX typesetting. Anyone taking up this challenge is invited to contact the author -- who would like to use such macros for definitive ASCII-Cyrillic documentation!

In the interim, one has a simple modus vivendi with essentially all TeX formats having 8-bit Cyrillic capablility -- one which requires no new macros at all! Namely, convert from ASCII-Cyrillic text to 8-bit Cyrillic text (with embedded TeX formatting), and then compose with TeX. (As will be explained, the TeX formatting commands are largely unchanged when expressed in ASCII-Cyrillic.) The converter "email-ru.tex" then serves as a preprocessor to TeX. One way to get good value from this approach is to break your TeX typescript into files that are either purely ASCII or else mostly Cyrillic. Only the latter sort undergo conversion. The two sorts of file can then be merged using TeX's `\input` command or LaTeX's `\include`.

Snag Warning

A few important TeX implementations, notably C TeX under unix, and a majority of implementations for the Macintosh OS, are currently unable to `\write` true octets > 127 --- as "email-ru.tex" requires in converting from ASCII-Cyrillic to 8-bit Cyrillic text. (This problem does not impact the conversion from 8-bit Cyrillic text to ASCII-Cyrillic.)

To solve this problem when it arises, the ASCII-Cyrillic package will rely on a tiny autonomous and portable utility "Kto8" that converts into genuine 8-bit text any text file which the few troublesome TeX installations may output.

The sign that you need to apply this utility is the appearance of many pairs ^^ of hat characters in the output of "email-ru.tex".

Ready-to-run binary versions of "Kto8" will progressively be provided for the linux, unix, Macintosh, and Windows operating systems. The most current distribution of "Kto8" is at http://topo.math.u-psud.fr/~lcs/Kto8/. See also the CTAN archive.

Crash course on Russian ASCII-Cyrillic

The 33 letters of the modern Russian alphabet, in alphabetic order, are typed:

```
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 w 'w q y h 'e 'u 'a
```

The corresponding Cyrillic glyphs are:

```
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
```

Similarly for capital letters:

```
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 W 'W Q Y H 'E 'U 'A
```

correspond to:

```
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
```

It is worth comparing this with the phonetic recitation of the alphabet (in an informal ASCII transcription):

```
ah beh veh geh deh yeh yo zheh zeh
ee (ee kratkoe) kah el em en oh peh
err ess teh oo eff kha tseh cheh
shah shchah (tv'ordyj znak) yerry
```

```
(m'agkij znak) (e oborotnoe) yoo ya
```

where parentheses surround descriptive names for letters that are more-or-less unpronouncable in isolation.

When there is a competing ergonomically "optimal" choice for typing a Russian character, the alternative may be admissible in ASCII-Cyrillic. Thus:

```
'g='z
's=w
c='t
'k=x
```

Incidentally, the strongest justification for typing "c" for a letter consistently pronounced "ts" is the traditional Russian recitation of the Latin (ASCII) alphabet:

```
ah beh tseh deh ...
```

For the Ukrainian Cyrillic "hard g" (not in the modern Russian alphabet), Russian ASCII-Cyrillic requires typing:

```
'{gup}
```

(and '{GUP} for the uppercase form). Similarly for other Cyrillic letters. The braces proclaim a Cyrillic letter and the notation is valid for every Cyrillic language.

For the Russian number character, which resembles in shape the pair "No", ASCII-Cyrillic uses the notation

```
'[No]
```

Similarly for the numerous other non-letters. Exceptionally, for this widely used symbol, the short form 'N is allowed. The square brackets proclaim a non-letter. One oddity to note is that for text double right quotes one types `'["]` (4 characters) and not `'['']` (5 characters) while for text double left quotes one types `'[``]` (5 characters) .

The two long notation schemes `'{...}` and `'[...]` afford a systematic way to represent all characters typed on any Cyrillic computer keyboard; and they leave room for future evolution.

The ASCII-Cyrillic expression for an octet >127 not encoded to any normalized character, is

```
!__xy
```

Here `__` is two ASCII underline characters, and xy is the two-digit lowercase hexadecimal representation of the octet. Imagine, for example, that, in the 8-bit Cyrillic text encoding, the octet number hex 8b (= decimal 139) is for non-text graphic purposes or else is undefined. In either case, it is rendered in conversion to ASCII-Cyrillic as

```
!__8b
```

Conversion from this back to the 8-bit form will work. However, although the 5 octet string "!__8b" is ASCII text, this text is not independent of 8-bit encoding. Thus, it is important to eliminate such "unencoded" or "meaningless" octets. A Cyrillic text file containing them is in some sense "illegal".

Tha ASCII letters are the English unaccented letters. The ASCII non-letter characters, namely:

```
! " # $ % & ' ( ) * + , - . /
0 1 2 3 4 5 6 7 8 9 : ; < = > ? @
[ \ ] ^ _ `
{ | } ~
```

are all common to Russian and English computer keyboards and 8-bit encodings. It is worth remembering these non-letters, since you can then identify ASCII text at sight. All of these, except on occasion `' ! \ ,` can be freely used in ASCII-Cyrillic typing of Russian prose; they are not altered under conversion to an 8-bit encoding.

ASCII-Cyrillic is not well designed for typing English sentences, but, since occasional English words or letters are used in Russian, ASCII-Cyrillic allows one to type, for example, `!U` for an isolated `U` and:

```
!Coca-!Cola          for        Coca-Cola
```

The special relationship with TeX

The converter "email-ru.tex" is programmed as a TeX macro package because TeX is perhaps the most widely and freely available utility that can do the job.

The relation with TeX runs deeper. TeX is a powerful stable and portable formatting system, and perhaps the most widely used system for scientific and technical documents. For a continental European language with an accented Latin alphabet (French for example), a TeX typescript is often created as an 8-bit text file that (just as for Russian) depends on 8-bit encoding. However TeX itself has always offered an alternative more prolix ASCII form for such accented letters. For example, `\'e` represents `e` with an acute accent. This ASCII form has always served to provide portable ASCII typescripts that are readable and editable. ASCII-Cyrillic seems to be the first ASCII scheme to offer something similar for all Russian TeX typescripts.

To let users type TeX commands with reasonable comfort in ASCII-Cyrillic, the latter preserves TeX control sequences like `\begin`. The familiar command

```
\begin{document}
```

is thus expressed as:

```
\begin{!document}
```

Russians use mostly ASCII letters in math mode. According to the usage of `!` already explained, the ASCII-Cyrillic $!e=!mc^2$ converts to Einstein's formula $e=mc^2$. The extra exclamation marks are an annoyance. However, for the same TeX output, you could type this formula without the extra exclamation marks --- provided you first define special `\mathcode` values for the octets of `\cyre`, `\crym`, `\cyrc`, etc. Consult the TeXbook about `\mathcode`.

The escape characters: The special roles played by the three characters `' ! \` impose a few strange rules in ASCII-Cyrillic typing. Notably, the ASCII prime `'` must sometimes be typed as `''` (two primes). Experimental use of "email-ru.tex" will allow the user to find his way as quickly as would detailed documentation. (Please report any needlessly complex or absurd behavior!)

Ukrainian ASCII-Cyrillic

This is similar to but distinct from the Russian mode and is not compatible with it.

The 33+1 letters of the modern Ukrainian alphabet, listed in alphabetic order are:

а б в г ґ д е є ж з и і ї й к л м
н о п р с т у ф х ц ч ш щ ю я ь ’

and the preferred Ukrainian ASCII-Cyrillic form is:

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

The 34th character is a Cyrillic apostrophe, a "modifier letter" that has various roles, among them those of the hard sign of Russian. The representation valid for all Cyrillic languages is '{apos}.

The phonetic recitation of this alphabet (using an informal ASCII transcription) is:

```
ah beh veh heh geh deh eh yeh
zheh zeh y(?) ee yee yot kah el
em en oh peh err ess teh oo eff kha tseh
cheh shah shchah yoo ya (m'akyj znak)
(apostrof)
```

The alternative short forms in Ukrainian ASCII-Cyrillic: are

```
h=g  's=w  c='t  'k=x
```

The following four letters do not occur in Russian:

```
ґ є і ї

<=>  '{gup} '{ie} '{ii} '{yi} (all Cyrillic languages)
<=>  'g  'e  i  'i   (short forms for Ukrainian)
<=>  (no Russian short forms)
```

Reciprocally, the following four Russian letters do not occur in Ukrainian:

```
ъ ы э ё

<=> '{hrdsn} '{ery} '{erev} '{yo}  (all Cyrillic)
<=> (no Ukrainian short forms)
```

```
<=> q  y  'e  'o    (short forms for Russian)
```

The following two letters are common to Ukrainian and Russian, but the ASCII-Cyrillic short forms are different.

```
    И  Ь

<=>  '{i}  '{sftsn}   (all Cyrillic)
<=>  y  q    (short forms for  Ukrainian)
<=>  i  h    (short forms for  Russian)
```

In Ukrainian ASCII-Cyrillic, the use of q as a short form for '{sftsn} is supported by the fact that the shape q rotated by 180 degrees is similar to that of '{sftsn} . But there is another reason for this choice. It permits one to use h as an alternative Ukrainian short form for '{g} --- which is natural since in many cases '{g} is pronounced like the harsh German h in "Horst".

Similarly for capital letters. In particular:

```
А Б В Г Ґ Д Е Є Ж З И І Ї Й К Л М
Н О П Р С Т У Ф Х Ц Ч Ш Щ Ю Я Ь '
```

have the Ukrainian ASCII-Cyrillic representation:

```
A  B  V  G  'G  D  E  'E  'Z  Z  Y  I  'I  J  K  L  M
N  O  P  R  S  T  U  F  X  'T  'C  W  'W  'U  'A  Q  '*
```

Long forms valid for all Cyrillic languages are:

```
'{A} '{B} '{V} '{G} '{GUP} '{D} '{E} '{IE} '{ZH} '{Z}
'{R} '{I} '{II} '{YI} '{J} '{K} '{L} '{M} '{N} '{O}
'{P} '{S} '{T} '{U} '{F} '{X} '{TS} '{CH}
'{SH} '{SHCH} '{YU} '{YA} '{SFTSN} '{APOS}
```

Note that the Ukrainian apostrophe '{APOS} is a **letter** and, unlike '{SFTSN}, it normally coincides with the lowercase version: normally '{APOS}='{apos}. In case of a distinction, '* will be '{apos}. Further, '{apos} normally has shape identical to the text right single quotation mark denoted in ASCII-Cyrillic by '[].

There is an official lossy ASCII transliteration for Ukrainian using the ligature concept, and it is supported by "email-ru.tex". See the Ukrainian national norm of 1996 summarized at:

```
http://www.rada.kiev.ua/translit.htm
```

Beware that the official Ukrainian transliterations of the six letters:

```
'{g}   '{ie}   '{yi}   '{ishrt}   '{yu}   '{ya}
```

are context dependent. This is a good reason for calling upon "email-ru.tex" to do the official transliteration.

The other aspects of ASCII-Cyrillic are the same for Ukrainian and Russian.

Vital statistics for ASCII-Cyrillic

ASCII-Cyrillic home page: (established December 2000)

```
topo.math.u-psud.fr/~lcs/ASCII-Cyrillic/ascii-cy.htm
```

ASCII-Cyrillic software directory:

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

Long term archiving: See the CTAN TeX Archive and its mirrors.

Copyright conditions: Gnu Public Licence.

Documentation: -currently included as ASCII text inside the converter "email-ru.tex".

Debts: The author owes many thanks, in particular:
-- to Stanislas Klimenko for an invitation to IHEP Protvino, Russia, in Fall 1977; ASCII-Cyrillic was conceived there;
-- to Irina Maxova'a for suggesting in November 1997 that the ASCII `w` represent the Cyrillic letter "sh" (`\cyrsh` in TeX);
-- to Gal'a Gor'a'cevskix for answering innumerable questions about Russian;
-- to the members of the Cyrillic TeX discussion list (CyrTeX-en@vsu.ru), moderated by Vladimir Volvovi'c, both for clarifying problems and for furnishing vital data. The

list archives are available at:

`https://info.vsu.ru/Lists/CyrTeX-en/List.html`

-- to Maksym Pol'akov (mpoliak@pcomp.nauu.kiev.ua) whose extensive advice was essential in establishing Ukrainian mode.

Date of most recent modification: July, 2001.

The author: (who welcomes comments)

Laurent Siebenmann
CNRS, Université de Paris-Sud
Orsay, France

lcs@math.u-psud.fr
lcs@math.polytechnique.fr
laurent@math.toronto.edu