

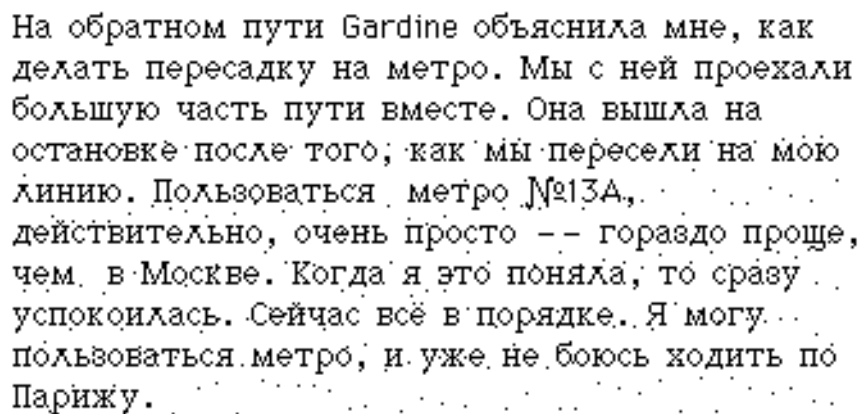
This is the ASCII-Cyrillic Home Page, PDF rendition.
N.B. The bitmaps probably look best at 100% size!

ASCII-Cyrillic and its converter email-ru.tex (beta version)

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 brief introduction is initially for Russian; but, further along, come the modifications needed to adapt to the Ukrainian alphabet.

Here is a fragment of Russian email. As far as the email system was concerned, the email message was roughly a sequence of "octets" or "bytes" (each 8 zeros or ones); where each octet corresponds to a character according to some 8-bit encoding. As originally typed and sent, it is probably readable (using a 8-bit Russian screen font) on most computers in any country where a Cyrillic alphabet is indigenous --- but rarely beyond.



На обратном пути Gardine 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 №13А,
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 по
Парижу.

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

The portability of 8-bit Cyrillic text is hampered by the frequent need to re-encode for another computer operating system. When the targeted encoding does not contain all the characters used, reencoding can become not just inconvenient but downright problematic.

The utility "email-ru.tex" converts this 8-bit text to and from ASCII-Cyrillic, the new 7-bit ASCII transcription of Russian. This scheme was designed to be both typeable and readable on every computer worldwide:

Na obratnom puti !Gardine 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, dejstvitel'no, ocen'no prosto -- gorazdo pro'we, cem v Moskve. Kogda 'a 'eto pon'ala, to srazu uspokoilash. Sejcas vs'o v por'adke. 'A mogu polhzovaths'a metro, i u'ze ne bo'ush xodith po Pari'zu.

Well chosen English (Latin) 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 enjoy a fixed meaning and shape governed by the universally used ASCII standard. Also, all 8-bit Cyrillic text encodings respect the ASCII standard where 7-bit characters are concerned.

In 7-bit ASCII-Cyrillic form, Russian prose is less than 4 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 4 percent in bulk drops to less than 1 percent when modern "gzip" compression is applied to both. Thus, there is virtually no penalty for storing Cyrillic text files in ASCII-Cyrillic form.

As the 7-bit ASCII-Cyrillic form can be converted by "email-ru.tex" back to any of the most used 8-bit encodings, one can also convert in 2 steps between 8-bit encodings.

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 Gardine 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 N013A, 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' po Parizhu.

Caveat: Accurate reconversion of existing ligature-based transcriptions back to 8-bit format 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 deleted.)

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

Beginner's operating instructions for using "email-ru.tex" as a converter are simple:-

- 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 onscreen by "email-ru.tex".

The most complete technical documentation for ASCII-Cyrillic is currently included *inside* the converter "email-ru.tex" in order to enhance the converter's autonomy. The present HTML format is probably more readable since Cyrillic character shapes are presented using universally valid GIF graphics. (Look also for a related PDF version.)

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 small 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 lunux, unix, Macintosh, and Windows operating systems. Here is the most current distribution of Kto8. See also the CTAN archive.

Quick Introduction to 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 Latin 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 = 'т
'k=x

Incidentally, the strongest justification for typing "c" for a letter consistently pronounced "ts" is the traditional Russian recitation of the Latin 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. The square brackets proclaim a non-letter. One oddity to note is '['] (not '['']) for text double right quotes.

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 that, in the 8-bit Cyrillic text encoding, the octet 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".

The ASCII non-letter characters are all common to Russian and English, namely:

! " # \$ % & ' () * + , - . /
0 1 2 3 4 5 6 7 8 9 : ; < = > ? @

[\] ^ - `
{ | } ~

It is worth remembering these, since you can then pretty well 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 occasional English words or letters are used in Russian, so ASCII-Cyrillic allows one to type !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 for e with an acute accent. This form has always served to provide exchangeable 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}

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 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 the alphabet (using an informal Latin 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} (for 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 Ъ

```
<=>  '{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 Г' Д Е Є Ж З И І І' Й К Л М
 Н О П Р С Т У Ф Х Ц Ч Ш Щ Ю Я Ъ ' ,

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 "Latin 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 transliterations of the six letters:

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


are **context dependent**. This is a good reason for relying on "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)

<http://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.

Link Index

- **Introduction to the converter email-ru.tex**
- **Quick Introduction to Russian ASCII-Cyrillic**
- **Quick Introduction to Ukrainian ASCII-Cyrillic**
- **Download ASCII-Cyrillic**
(<http://topo.math.u-psud.fr/~lcs/ASCII-Cyrillic/>)
- **Download Kto8** (<http://topo.math.u-psud.fr/~lcs/Kto8/>)

The author (who welcomes comments):

Laurent Siebenmann

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

Thanks!

Many thanks are owed to the members of the Cyrillic TeX discussion list (CyrTeX-en@vsu.ru) for both clarifying the problems that a utility such as this one should address and furnishing vital data. The list archives are available at:

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

Advice from Maksym Polyakov (mpoliak@pcomp.nauu.kiev.ua) was essential in establishing the Ukrainian mode.

Date last modified: 6 February, 2001.