Cyrillic languages support in LaTeX

Vladimir Volovich, Werner Lemberg and LaTeX3 Project Team.

12 March 1999

Contents

Abstract

This document contains basic information on the Cyrillic setup for LaTeX: how to get the fonts, how to set them up, how to use the interface, its interaction with babel, etc. This is only a first draft of the document and it will probably be modified in future; so please send in comments on it via the `latexbug` system (see below).

1 Introduction

Most Latin-based European languages were supported in LaTeX by introducing the T1 font encoding and by using the fontenc and inputenc packages; these use only standard TeX means to support any 8-bit input encoding and this one standard font encoding. The restriction to a single font encoding guarantees that multiple languages can happily coexist in one document (e.g., hyphenation will be correct for all languages).

Starting with the December 1998 Release, LaTeX finally supports Cyrillic languages. This support is based on the new standard Cyrillic TeX font encodings—T2A, T2B, T2C, and X2. The first three of these satisfy some basic requirements for LaTeX T* encodings, and thus can be used in multi-lingual documents with other languages based on standard font encodings.

The reason why we need four different Cyrillic font encodings is that these font encodings support *all* the Cyrillic languages that have been used during the twentieth century (see Section 4)! The number of Cyrillic glyphs is large, so they cannot be represented with 128 character slots; the other (lower) 128 slots are reserved for Latin letters and other invariant symbols that are needed for the encoding to be a conformant LaTeX T encoding.

There are some glyphs in the T2* encodings which do not yet have associated characters in *Unicode*, the world-wide character standard. Also, one more font encoding, T2D, is planned for a forthcoming release of LaTeX. A lot of Cyrillic input encodings are already supported (see Section 5), and additional encodings could be added easily.

1.1 Acknowledgments

The work on T2* encodings was carried out by the T2 Team, led by Alexander Berdnikov (other members are Mikhail Kolodin and Andrew Janishewskii). The LH fonts were produced by Olga Lapko (with A. Khodulev). The T2 bundle and ruhyphen package were written by Werner Lemberg and Vladimir Volovich (except that the concrete hyphenation patterns which are part of ruhyphen came from individual authors). The support for the Ukrainian language was prepared by Andrij Shvaika.

2 Installation

The fontenc and inputenc packages are installed automatically in every base LaTeX distribution.

All the necessary extra files to use with these packages for Cyrillic are in the cyrillic bundle, which at present contains the following: four font encoding definition files (`t2aenc.def`, `t2benc.def`, `t2cenc.def`, `x2enc.def`); several input encoding definition files (all the other `*.def` files), and font definition files (`*.fd`). The installation of these is described here.

2.1 Fonts

The default font families in LaTeX are the Computer Modern families, namely the CM fonts (`OT1` encoded) and the EC fonts (`T1` encoded). The LH fonts, which are now available, provide Computer Modern fonts for all Cyrillic font encodings. They are designed to be compatible with the EC fonts, and they provide the same font shapes and sizes; they are available at `CTAN:fonts/cyrillic/lh` (the latest version is 3.20). The installation instructions for the fonts are in the file `INSTALL` in the font distribution.

Other fonts, including Type 1 fonts, can also be used, provided that their encoding (for TeX) is `T2`-compatible. Some ready-to-use packages supporting such fonts are also available, e.g., at `ftp://ftp.vsu.ru/pub/tex` (they should soon be on CTAN). Currently, you will find two packages there: PsCyr, which contains some freely distributable Cyrillic Type 1 fonts with support for LaTeX; and c1fonts, which contains virtual fonts similar to the AE fonts package using the BlueSky and BaKoMa fonts available from CTAN (see the `README` file in that package for detailed information). Further font packages are expected soon.

2.2 Hyphenation patterns

You can find a collection of hyphenation patterns for the Russian language in the ruhyphen package at `CTAN:language/hyphenation/ruhyphen`. These patterns support the `T2*` encodings, as well as other popular font encodings used for Russian typesetting (including the Omega internal encoding). Patterns for other Cyrillic languages should be adapted to work with the `T2*` encodings.

2.3 babel support for Russian and Ukrainian

Version 3.6k of babel includes support for the `T2*` encodings and for typesetting both Russian and Ukrainian texts using the Cyrillic letters. The temporary fontencoding `LWN`, which was used in earlier releases of babel, will be withdrawn in the near future and replaced by the `OT2` encoding.

2.4 Getting pre-built packages

Many of the major TeX distributions, such as teTeX, fpTeX and TeXlive, contain (or soon will) everything that is needed, including the LH fonts, ruhyphen and the latest version of babel. We hope that all TeX distributions will soon include all of these, so that the chances are that you will not need to install this by yourself (but it is not difficult).

If you are using emTEX, MikTEX, or fpTEX, you can download the ruemtex package from `ftp://ftp.vsu.ru/pub/tex`.

3 Usage

Support for Cyrillic is based on these standard LaTeX mechanisms: the fontenc and inputenc packages (and on babel). Thus the basic principles for its use are similar to those for other European languages: you simply add, to your document preamble, lines like the following.

```
\usepackage[T2A]{fontenc}
\usepackage[koi8-r]{inputenc}
```

Here you can put any desired input encoding instead of `koi8-r`: for example, it would be `cp866` if you are using a MS-DOS text editor with this Cyrillic code page to prepare your documents, or `cp1251` if you are a MS Windows user with Cyrillic support. A full list of the available Cyrillic encodings can be found in Section 5 and in the file `cyinpenc.dtx`.

Documents are, naturally, not restricted to a single font encoding; this is essential for multi-lingual journals or documents. Such changes can be made by using the `\fontencoding` command as part of a font-change. However, it is best to access these font encodings via a higher-level interface.

Since such changes are often closely related to other language-dependent settings, it is often sensible to use the babel system, which provides further useful 'localisation' and standardised multi-lingual interfaces (for further details, see Section 2.3). Then you can use lines like the following in your document:

```
\usepackage[koi8-r]{inputenc}
\usepackage[russian]{babel}
```

This will automatically choose the default font encoding for Russian, which is `T2A`, if available. Documentation of the complete set of font-encoding selection rules can be found in `cyrillic.dtx` which is part of `rusbabel`.

These LaTeX interfaces are very convenient because they make your documents completely portable, being based solely on standard TEX features. This will mean that your documents can be processed on any TEX system without any need for re-encoding to the 'native' encoding used on each platform; this is because the encoding of the document is specified in the document itself.

Moreover, if necessary, more than one input encoding can be used within a document; this could be useful if, for example, you need to combine articles prepared by authors on different machines. Each part of the document is then identified by a `\inputencoding` command, which can therefore only be used between paragraphs.

Please note that you must always use the two standard LaTeX commands, `\MakeUppercase` and `\MakeLowercase` to produce uppercase or lowercase text in your documents. This is because `\uppercase` and `\lowercase` will not work at all for Cyrillic (note that these latter two commands are not, and never have been, available for use directly in LaTeX documents).

4 Font encodings for Cyrillic languages

The Cyrillic font encodings support the following languages. Note that some languages can be properly typeset with more than one encoding.

`T2A`: Abaza, Avar, Agul, Adyghei, Azerbaijani, Altai, Balkar, Bashkir, Bulgarian, Buryat, Byelorussian, Gagauz, Dargin, Dungan, Ingush, Kabardino-Cherkess, Kazakh, Kalmyk, Karakalpak, Karachaevskii, Karelian, Kirghiz, Komi-Zyrian, Komi-Permyak, Kumyk, Lak, Lezghin, Macedonian, Mari-Mountain, Mari-Valley, Moldavian, Mongolian, Mordvin-Moksha, Mordvin-Erzya, Nogai, Oroch, Osetin, Russian, Rutul, Serbian, Tabasaran, Tadzhik, Tatar, Tati, Teleut, Tofalar, Tuva, Turkmen, Udmurt, Uzbek, Ukrainian, Hanty-Obskii, Hanty-Surgut, Gipsi, Chechen, Chuvash, Crimean-Tatar.

`T2B`: Abaza, Avar, Agul, Adyghei, Aleut, Altai, Balkar, Byelorussian, Bulgarian, Buryat, Gagauz, Dargin, Dolgan, Dungan, Ingush, Itelmen, Kabardino-Cherkess, Kalmyk, Karakalpak, Karachaevskii, Karelian, Ketskii, Kirghiz, Komi-Zyrian, Komi-Permyak, Koryak, Kumyk, Kurdian, Lak, Lezghin, Mansi, Mari-Valley, Moldavian, Mongolian, Mordvin-Moksha, Mordvin-Erzya, Nanai, Nganasan, Negidal, Nenets, Nivh, Nogai, Oroch, Russian, Rutul, Selkup, Tabasaran, Tadzhik, Tatar, Tati, Teleut, Tofalar, Tuva, Turkmen, Udyghei, Uigur, Ulch, Khakass, Hanty-Vahovskii, Hanty-Kazymskii, Hanty-Obskii, Hanty-Surgut, Hanty-Shurysharskii, Gipsi, Chechen, Chukcha, Shor, Evenk, Even, Enets, Eskimo, Yukagir, Crimean Tatar, Yakut.

`T2C`: Abkhazian, Bulgarian, Gagauz, Karelian, Komi-Zyrian, Komi-Permyak, Kumyk, Mansi, Moldavian, Mordvin-Moksha, Mordvin-Erzya, Nanai, Orok (Uilta), Negidal, Nogai, Oroch, Russian, Saam, Old-Bulgarian, Old-Russian, Tati, Teleut, Hanty-Obskii, Hanty-Surgut, Evenk, Crimean Tatar.

The `X2` encoding was designed to support all the above languages. Its name does not start with `T` because, for example, it contains no Latin letters (it is purely a Cyrillic glyph container); it therefore cannot be used in mixed-script documents along with the other `T*` encodings. Please consult Section 6.4 *Naming conventions* of the file `fntguide.tex` in the base LaTeX distribution for details of the differences between LaTeX font encodings and how they are named.

There are two other LaTeX Cyrillic font encodings, `OT2` and `LCY`, that are not included in the base LaTeX distribution. The first is a 7-bit encoding (hence the `O`) developed by the AMS; it is useful for typesetting relatively small fragments of text in Cyrillic, using a Latin transliteration scheme. The other, `LCY`, is an 8-bit Cyrillic encoding which is not compatible with the requirements for LaTeX `T*` encodings (hence the `L`); thus it is not suitable for typesetting multi-lingual documents, but it can be used in Plain TeX-based macro packages because it is an extension of `OT1`. These two encodings are supported by babel and by ot2cyr.

5 Input encodings

Several Cyrillic code-pages are widely used. Currently, LaTeX contains support for 20 Cyrillic input encodings (some of which are variants of each other).

- `cp855` — the standard MS-DOS Cyrillic code-page.
- `cp866` — the standard MS-DOS Russian code-page. Several code-pages very similar to this are also supported (the differences are all in the range 242–254).
  - `cp866av` – the 'Cyrillic Alternative' code-page (an alternative variant of cp866);
  - `cp866mav` – the 'Modified Alternative Variant';
  - `cp866nav` – the 'New Alternative Variant';
  - `cp866tat` – an experimental Tatarian code-page.
- `cp1251` — the standard MS Windows Cyrillic code-page.
- `koi8-r` — a standard Cyrillic code-page widely used in UNIX-like systems for Russian language support that is specified in RFC 1489. The situation with `koi8-r` is somewhat similar to that for `cp866`: there are several similar code-pages which coincide for all Russian letters but add some other Cyrillic letters. The following are supported:
  - `koi8-u` – for Ukrainian;
  - `koi8-ru` – this is described in a draft RFC document specifying a widely used character set for mail and news exchange in the Ukrainian internet community, as well as for presenting WWW information resources in the Ukrainian language;

  - `isoir111` – the ISO-IR-111 ECMA Cyrillic Code Page.

- `iso88595` — the ISO 8859-5 Cyrillic code-page (also called ISO-IR-144).
- `maccyr` — the Apple Macintosh Cyrillic code-page (also known as Microsoft cp10007) and `macukr`, the Apple Macintosh Ukrainian code-page, very similar to the Cyrillic code-page.
- The Mongolian code-pages: `ctt dbk mnk mos ncc mls`. These code-pages were taken from Oliver Corff's 'MonTEX' package (available at `CTAN:language/mongolian/montex`). Since the `T2*` encodings support the Mongolian Cyrillic script, it is convenient to have support for Mongolian input encodings as well. Pointers to documentation for these code-pages will be much appreciated.

6 Reporting bugs

In case you find a bug and want to report it, please follow the guidelines given in the file `bugs.txt` in the base LATEX distributions. Note that there is a category specifically for reporting any bugs that occur only when using Cyrillic fonts or support packages.

7 Miscellanea in the T2 bundle

The T2 bundle at `CTAN:macros/latex/contrib/supported/t2` contains some other useful files, including support for Plain TEX-based macro packages, support for BibTEX and MakeIndex (see also the xindy program and package—highly recommended for making indices with Cyrillic), support for the fontinst package, mapping tables relating these Cyrillic font encodings (and input encodings) to the Unicode character names and slots (these are in the subdirectory `enc-maps`), and more!

To produce documented source listings of the T2 package, run LATEX on the `*.dtx` and `*.fdd` files therein.

When typesetting Cyrillic texts, there is a tradition of using Cyrillic letters (in some situations) within math formulæ, in exactly the same way as most of the world uses Latin letters. By default this does not work, because symbols declared with `\DeclareTextSymbol` may not be used in math.

If you need within math to 'transparently' typeset glyphs declared in font encoding definition files, then you could try using the experimental mathtext package, which is also in the T2 bundle. Note that this package uses up at least one additional math alphabet per font encoding. For this and other reasons, The LATEX3 Project Team considers that this experimental extension to LATEX's glyph-handling mechanisms should be used with caution; but please try it out and send us your opinions and ideas. Note that it is not included in the core of LaTeX because both the coding and the interfaces are likely to change at some point in the future.

Finally, here are some pointers to further information:

```
http://www.cemi.rssi.ru/cyrtug
http://xtalk.price.ru/tex
```