Typesetting multilingual verbatim text with pdflatex

Nelson H. F. Beebe

Department of Mathematics
University of Utah
Salt Lake City, UT 84112, USA

18 December 2019

Abstract

Producing multilingual verbatim displays in pdflatex from Unicode UTF-8 input is more difficult than expected, due to its history of 8-bit input encoding. This document shows how some of those limitations may be overcome, making it relatively easy to document computer software examples for programming languages that support Unicode strings.

1 Introduction

For technical documents about modern programming languages that support Unicode UTF-8 character strings, I need to typeset computer input and output in multiple character sets using pdflatex.

Both lualatex and xelatex are fully adapted to Unicode input, and native Unicode fonts. However, to maintain typographical compatibility with earlier versions of my documents, I require pdflatex, which is based on 8-bit characters with 256 glyphs in a font.

The babel package allows specification of the languages needed in a document, and its `\selectlanguage {...}` commands allow you to switch among them in your prose. Within a language text block, hyphenation patterns, spacing between words and sentences, and spacing around punctuation, depend on the current language.

However, the LaTeX `verbatim` environment, and the associated inline `\verb|...|` command, are highly specialized: they cause the TeX engine to enter a special mode where almost all characters, including spaces and newlines, are typeset as themselves, and that mode is only left when the closing delimiter string is reached. In particular, it is not possible to switch fonts during verbatim processing. Instead, the verbatim font at entry remains in effect, and is normally a fixed-width (typewriter) font. It can be redefined with a command like this: `\renewcommand {\ttdefault} {cmtt}`.

With a bit of essential setup in the LaTeX document preamble, it is easy to typeset a mixture of English and Russian in verbatim mode. For example, with `pdflatex` from TeX Live 2018 or later, the Unicode UTF-8 input stream

```
\usepackage [russian,english] {babel}
...
%% Change the body font to one with a large Unicode repertoire:
\renewcommand {\rmdefault} {NotoSerif-TLF}
...
%% Redefine \ttdefault so that both standard verbatim and fancy
%% Verbatim use the same font with a large Unicode repertoire:
\renewcommand {\ttdefault} {NotoSansMono-TLF}
...
%% Redefine the LaTeX internal verbatim font selection to
%% use the T2A encoding for Latin + Cyrillic input:
\makeatletter
    \renewcommand {\verbatim@font}
                  {%
                      \fontencoding {T2A}%
                      \fontfamily   {\ttdefault}%
                      \selectfont
                  }
\makeatother
...
\begin{document}
...
\selectlanguage {russian}%
\begin{verbatim}
  War and peace / Война и мир / Voyna i mir
\end{verbatim}
\selectlanguage {english}
```

produces this output:

```
War and peace / Война и мир / Voyna i mir
```

That works because a 256-character Russian typewriter font has standard ASCII characters in the lower 128 slots, with Cyrillic letters in the upper 128 slots, so both Latin and Cyrillic text can be used together. That design is likely universal in Cyrillic fonts, because literary, newspaper, and technical texts in the many languages that use the extended Cyrillic alphabet sometimes use phrases in the Latin alphabet. Bibliographic data and reference lists are also likely to require both alphabets.

If we mix additional alphabets into our English and Russian verbatim text, the complication for `pdflatex` is that the current font encoding only contains 256 slots, and there is no space for more letters. Thus, the way to get Greek text into mixed language LaTeX `verbatim` displays is to typeset it *outside* those displays, save the result in a LaTeX named box, and then use the capability of the `fancyvrb` package to embed commands in the `Verbatim` environment to insert the named box.

A box with Greek text is created with this input:

```
\newsavebox {\greekgreeting}
\savebox    {\greekgreeting}%
            {%
               \mbox {%
                           \selectlanguage {greek}%
                           \greekcolor
                           \texttt         {Greek text here}%
                     }%
            }
```

The same technique can be used for alphabets other than Greek, provided that we have font support for them, and that we take care to change input encodings with suitable `\fontencoding{...}` and `\selectlanguage {...}` commands. For example, we can create a named box in Icelandic with

```
\newsavebox {\icelandicgreeting}
\savebox    {\icelandicgreeting}%
            {%
               \mbox {%
                           \selectlanguage {icelandic}%
                           \icelandiccolor
```

```
                \fontencoding   {T1}%
                \texttt         {Icelandic text here}%
            }%
        }
```

and then typeset it in a fancy `Verbatim` display with

```
\selectlanguage {russian}
\begin{Verbatim}[commandchars=|\[\]]
  Hello, good day! / Здравствуйте! /
  |usebox[|greekgreeting] / |usebox[|icelandicgreeting]
\end{Verbatim}
\selectlanguage {english}
```

to produce this output in four languages:

```
Hello, good day! / Здравствуйте! /
Χαίρετε, καλή μέρα! / Halló, góður dagur!
```

Because the scope of the `\selectlanguage{...}` command does not extend beyond its current TEX group, we do not need `\selectlanguage {english}` inside the `\mbox{...}` commands to revert to the default document language.

2 Font style variations in Greek

Here is how the `\textxx{...}` font wrappers affect Greek text:

Greek bold	**Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ**
Greek italic	*Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ*
Greek medium	Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ
Greek normal	Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ
Greek roman	Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ
Greek sans serif	Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ
Greek slanted	*Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ*
Greek small caps	Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ
Greek typewriter	`Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ`
Greek typewriter bold	**`Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ`**
Greek typewriter italic	`Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ`
Greek upright	Χαίρετε, καλή μέρα = ΧΑΪΡΕΤΕ, ΚΑΛΉ ΜΈΡΑ
Latin transliteration	`Chaírete: kalí méra`

The variations available depend on the font family selected for `\rmdefault`. Our choice of *NotoSerif-TLF* offers a larger selection than is available with many older font families.

3 Font style variations in Russian

Here is how the `\textxx{...}` font wrappers affect Russian text:

Russian bold	**Добрый день! = ДОБРЫЙ ДЕНЬ!**
Russian italic	*Добрый день! = ДОБРЫЙ ДЕНЬ!*
Russian medium	Добрый день! = ДОБРЫЙ ДЕНЬ!
Russian normal	Добрый день! = ДОБРЫЙ ДЕНЬ!
Russian roman	Добрый день! = ДОБРЫЙ ДЕНЬ!
Russian sans serif	Добрый день! = ДОБРЫЙ ДЕНЬ!
Russian slanted	*Добрый день! = ДОБРЫЙ ДЕНЬ!*
Russian small caps	Добрый день! = ДОБРЫЙ ДЕНЬ!
Russian typewriter	`Добрый день! = ДОБРЫЙ ДЕНЬ!`
Russian typewriter bold	**`Добрый день! = ДОБРЫЙ ДЕНЬ!`**
Russian typewriter italic	`Добрый день! = ДОБРЫЙ ДЕНЬ!`
Russian upright	Добрый день! = ДОБРЫЙ ДЕНЬ!
Latin transliteration	`Dobryj den'`

4 Examples of multilingual verbatim displays

This is literal roman text in standard `texttt` mode in NotoSansMono-TLF:

```
abcde...vwxyz
```

This is literal Cyrillic text in standard `verbatim` mode:

```
Здравствуйте
```

This is literal roman and literal Cyrillic text in standard `verbatim` mode:

```
Hello! Здравствуйте!
```

This is literal Greek typewriter text, typeset inside a `\texttt{...}` macro: Χαίρετε, καλή μέρα

The input for that example looks like this:

```
..., typeset inside a \verb=\texttt{...}= macro:
%
\selectlanguage {greek}%
    \texttt{Greek text here}%
\selectlanguage {english}
```

This is literal English, literal Russian, and named-box Greek text in fancy `Verbatim` mode:

```
Hello! /  Здравствуйте! / Χαίρετε, καλή μέρα!
```

The input for that example looks like this:

```
\selectlanguage {russian}%
\begin{Verbatim}[commandchars=|\[\]]
    Hello! /  Здравствуйте! /  |usebox[|greekgreeting]
\end{Verbatim}
\selectlanguage {english}
```

This is literal English and literal Russian text in standard `verbatim` mode:

```
Hello! Здравствуйте!
```

This is named-box Greek text in prose: Χαίρετε, καλή μέρα!.

This is named-box Greek text in fancy `Verbatim` mode:

```
Χαίρετε, καλή μέρα!
```

This is literal English and named-box Greek text in fancy `Verbatim` mode:

```
Hello, good day /  Χαίρετε, καλή μέρα!
```

This is literal English, named-box Russian, and named-box Greek text in fancy `Verbatim` mode:

```
Hello, good day / Здравствуйте, добрый день! / Χαίρετε, καλή μέρα!
```