language and fonts

TEX in Polish

Erik Frambach
Faculty of Economics
University of Groningen
E.H.M.Frambach@eco.rug.nl

abstract

Writing in Polish with TEX requires a few tricks. In Polish you need several accents that are often not available in 'standard' fonts. Some TEX macros can solve this problem more or less. We will show the pros and cons.

Another 'problem' is input encoding. One can use 8-bit input in combination with the corresponding codepage definition, or a 7-bit encoding with a few extras to make typing easier. Both methods will be discussed.

This article reflects the content of a lecture held at the NTG meeting on 11 November 1999.

keywords

Polish, ogonek, input encoding

Dzień dobry!

Any idea what the following text means?

> Serdecznie dziękuje nie ma za co, proszę. Przepraszam, do widzenia do zobaczenia na razie nie rozumiem. Jak to się, mówi po polsku? Jak masz na imię jak się pani nazywa miło mi cię. Pana panią poznać jak się masz jak się pan ma dobrze. Źle tak sobie jako tako gdzie jest toaleta?

Actually, it's rubbish. I just combined some standard phrases into something that *looks* like a plain paragraph. But, anyway, the listing below will give you some idea of the meaning of some phrases I abused.

□ Dzień dobry = Good day
□ Serdecznie dziękuje = Thank you
□ Nie ma za co, Proszę = You're welcome
□ Przepraszam = Excuse me
□ Do widzenia; do zobaczenia = See you later, goodbye
□ Na razie = Farewell, bye
□ Nie rozumiem = I don't understand
□ Jak to się, mówi po polsku? = How do you say this in Polish?
□ Jak masz na imię? Jak sie, Pani nazywa? = What's your name?
□ Miło mi cię (pana, panią) poznać = Nice meeting you
□ Jak się, masz? Jak się pan ma? = How are you?
□ Dobrze = Good
□ Źle = Bad
□ Tak sobie; Jako tako = Alright
□ Gdzie jest toaleta? = Where is the toilet?

Polish problems

When writing Polish text in TEX we have to deal with the following problems:

□ TEX accents:
The *acute* accent is used a lot: e.g. ń, ś, ć, ó. The *dot* accent is also used in Polish: ż. Both are readily available in all fonts and can be accessed using the standard TEX commands `\'` and `\.` respectively. These accents are acceptable but could be improved. E.g., the dot accent may seem too thick and the `\'` accent should be flattened slightly when applied to capitals. There is a Polish version of the Computer Modern fonts that implements these changes. Instead of `cm*.*` they are called `pl*.*` and they are fully compatible with the Computer Modern fonts. Both PL fonts and EC fonts use 'Cork encoding' so they are compatible with respect to Polish diacritics.

□ Non-TEX accents:
The *ogonek* accent is used on several characters. It's the small 'tail' in e.g. ę and should not be confused with the *cedilla* that looks like this: ȩ. The ogonek is more difficult to typeset because many fonts do not support it. So we have to either make a kludge or extend the fonts we use with extra characters. LATEX users can use the `ogonek` package. It defines a macro `\k` that will emulate an ogonek on its parameter. The results, however, are not optimal, to say the least. A better solution is to use 'Polished' fonts. The EC fonts also contain true ogonek characters. Below is a comparison of Computer Modern using the LATEX `ogonek` package, the Polish version of the Computer Modern fonts, and Times Roman using the LATEX `ogonek` package and the real ogonek accent provided in the font:

Computer Modern (ogonek): ęĄą

Computer Modern (Polish): ęĄą

Times Roman (ogonek): ęĄą

Times Roman (real): ęĄą

□ Special characters:
In Polish you need the 'l-slash' and the 'L-slash'. Most fonts have these characters, and they can be accessed by the standard TEX commands `\l` and `\L`. However, the CM versions of these character are hardly acceptable in Polish. In the Polish versions the angle and the length of the bar are different:

CM Roman: łŁ

PL roman: łŁ

□ Hyphenation:
Because of the many accents in Polish texts hyphenation can be a problem. If you simply use standard TEX commands for accents hyphenation will not work properly. That is, hyphenation will not be incorrect but many valid hyphenation points will not be considered. 8-bit encoded hyphenation patterns in combination with 8-bit input (or emulated 8-bit input) can solve this problem.

Polish notation

The Polish are already famous for their 'reverse Polish notation' so why not build another one. Instead of using backslashes it's more convenient to use forward slashes for writing Polish text (you could say the backslash is 'reversed'!).

In Polish formats the slash is actually a character that can be 'activated' and 'deactivated' using `\prefixing` and `\nonprefixing`. When active the slash will take care of applying the correct accent to the following character. In Polish it's always clear which one it should be, unlike in e.g. French or Dutch. So, `/s` will produce ś, `/z` will produce ż and `/e` will produce ę. This method works for Polish (and not for e.g. French) because in Polish there can be no confusion over which accent should be used. An 'e' can only become ę, not e.g. ë, etc. There is just one exception to the rule: ź. This one can be coded as `/x`. Uppercase variants also exist. Naturally the slash macro will also make sure that hyphenation works properly. Very neat.

Since the slash notation is so powerful, why not use it in hyphenation patters as well? Here is a snippet of the Polish hyphenation patterns that shows how the slash is made active and defined as a macro that outputs an accented character based on its parameter.

```
\catcode'\/=13
\def/#1{%
  \ifx#1a^^a1\else
  \ifx#1c^^a2\else
  \ifx#1e^^a6\else
  \ifx#1l^^aa\else
  \ifx#1n^^ab\else
  \ifx#1o^^f3\else
  \ifx#1s^^b1\else
  \ifx#1x^^b9\else
  \ifx#1z^^bb%
  \fi\fi\fi\fi\fi\fi\fi\fi\fi}%
```

Lowercase codes are provided in order to make the hyphenation patters work for these accented characters:

```
\lccode "A1 = "A1 % /a (161)
\lccode "A2 = "A2 % /c (162)
\lccode "A6 = "A6 % /e (166)
\lccode "AA = "AA % /l (170)
\lccode "AB = "AB % /n (171)
\lccode "F3 = "F3 % /o (243)
\lccode "B1 = "B1 % /s (177)
\lccode "B9 = "B9 % /x (185)
\lccode "BB = "BB % /z (187)
```

Now the slash notation can be used in hyphenation patterns:

```
\patterns{
./c/c8
./c/l8
./c/n8
./c/s8
./c/z8
./c8
./cb8
.b/c8
.b/l8
.b/n8
.b/s8
.bc8
.bd8
.be2z3
.be3z4an
.ca/lo3/s2
.ca/lo3k2
.nad/srod5ziem
.zado/s/cu4
.zado2/s/c3
po/lu3d2ni
```

```
sze4/s/c
...
```

Input encoding

For Polish people using MS-Windows it's natural to use codepage 1250 (East European) to input text because it supports all the characters they need. On Unix ISO8859-2 is used, and on MS-DOS and Atari it's usually codepage 852. On Amiga and Macintosh machines you may find still other encodings! However, the LaTeX `inputenc` package can be used to process such text without change. You only need to include the following statement in the preamble of the LaTeX document:

```
\usepackage[cp1250]{inputenc}
```

The `inputenc` package will map special characters to the appropriate macros. Below is a small part of `cp1250.def` that shows how this is done. Note how the `\k` macro is used for ogonek.

```
\DeclareInputText{156}{\@tabacckludge's}
\DeclareInputText{163}{\L}
\DeclareInputText{165}{\k A}
\DeclareInputText{175}{\.Z}
\DeclareInputText{179}{\l}
\DeclareInputText{185}{\k a}
\DeclareInputText{186}{\c s}
```

Encoding translation tables

Another way of dealing with the East European codepage is to use an *encoding translation table*, a feature supported by some TeX implementations such as emTeX and Web2c (teTeX, fpTeX) and MikTeX. In that case the input is translated from one encoding into another one *before* TeX reads it. This can be very convenient but one must be aware that files that depend on this feature may not be processed correctly on other TeX systems. Below is a snippet of such a translation file:

```
%% cp1250pl.tcx:
%%   encoding translation table for TeX
%% source (TeX input):
%%   cp1250 (Windows East European)
%% target (TeX intestines):
%%   PL and QX encoding (Polish PL and QX fonts)
%
%% MAIN ENCODING TABLE:
...
0x9c 0xb1 % 156 177 sacute
0x9e 0xba % 158 186 zcaron
0x9f 0xb9 % 159 185 zacute
0xa3 0x8a % 163 138 Lslash
0xa5 0x81 % 165 129 Aogonek
...
```

An encoding translation table can used in two ways. The TeX compiler can be invoked with a parameter like this:

```
tex -translate-file=cp1250pl myfile
```

You can also add this parameter to the file itself. It must be a *comment* on the first line like this:

```
%& -translate-file=cp1250pl
```

Integration

Now that we have solutions to the basic problems for Polish TeX we can integrate them to make (TeX) life easier.

- The Polish have defined their own variant of plainTeX called MeX. It does everything that plain TeX does but also supports the slash notation and uses PL fonts. It also gives access to French style guillemots used in nested citations, and it supports repetition of the hyphen character on the beginning of the next line when breaking a paragraph into lines.
- A restricted LaTeX format was made in which only US language and Polish language is supported, without using Babel: it uses `polski.sty`. Naturally the PL fonts are used by default and EC fonts are supported using `\usepackage[T1]{fontenc}`. The `polski` package can also used with standard LaTeX if you want to use the Polish hyphenation patterns.
- A Polish version of the MakeIndex program was produced because the accented characters require a more complex sorting algorithm. This version can be adopted to other 8-bit encodings as well.
- An 8-bit version of the BibTeX program was produced. It knows how to deal with the Polish accented characters.

References

A detailed report called 'Polishing TeX: from ready to use to handy in use' by Bogusław Jackowski and Marek Ryćko was presented at the EuroTeX'92 conference in Prague. Staszek Wawrykiewicz made various suggestions to improve this article.

Do widzenia!

Goodbye!