TEX in Czech environment

Jaroslav Šnajdr
University of West Bohemia
Plzeň, Czech Republic
snajdr5@students.zcu.cz

Abstract

History of Czech typography, Czech alphabet, typefaces of Czech authors, typesetting traditions. CS-fonts, details of MF implementation, Type1 version, comparison with EC fonts. Czech in plain and LATEX: various problems and solutions – hyphenation, font encodings, special macros, future directions.

Introduction

Czech language uses Latin alphabet, where the meaning of many letters is altered by diacritical marks. Accented letters were introduced by Mistr Jan Hus at the beginning of the 15th century. Before his orthographical reform Czech was written using digraphs. In today's Czech, digraphs are not used at all (on the contrary, Polish still has "sz", "cz" and others) which causes particular abundance of diacritical marks in Czech texts.

Unfortunately, at the time of Hus's reform the only technology of book production was manual writing, which is the only technology, where are no problems with the diacritical marks — in metal typesetting, it is difficult to handle all those tiny gadgets and in computer area, it is difficult to handle the mess of various font encoding and badly designed accents in typefaces, that are mostly developed in English environment. Even the best designed accents bring irregularities into the grayness of the page. On figure 1, you can see one page of the first Czech book, "Kronika trojánská" (Trojan chronicle). It was printed in Plzeň in 1460 and you can see that it uses both digraphs and accented letters (unlike today's Czech).

In the sequel, we shall describe all kinds of Czech accents and explain, how they should look in a well designed typeface. As an example, you can see on figures 2 and 3 two beautiful Czech typefaces "Manuscript" and "Figural" made by one of the best Czech typographers, Oldřich Menhart. He is also the author of many very nice textbooks on typeface design — they were a good source of information for this article.

Accents

Accents on uppercase letters are usually lower and flatter than those on lowercase. In almost every PostScript font made abroad accented characters are supposed to be composed from letter and separate accent, but the accent is present in only one form in the font. This causes ugly uppercases and problems in typesetting with small baselineskip.

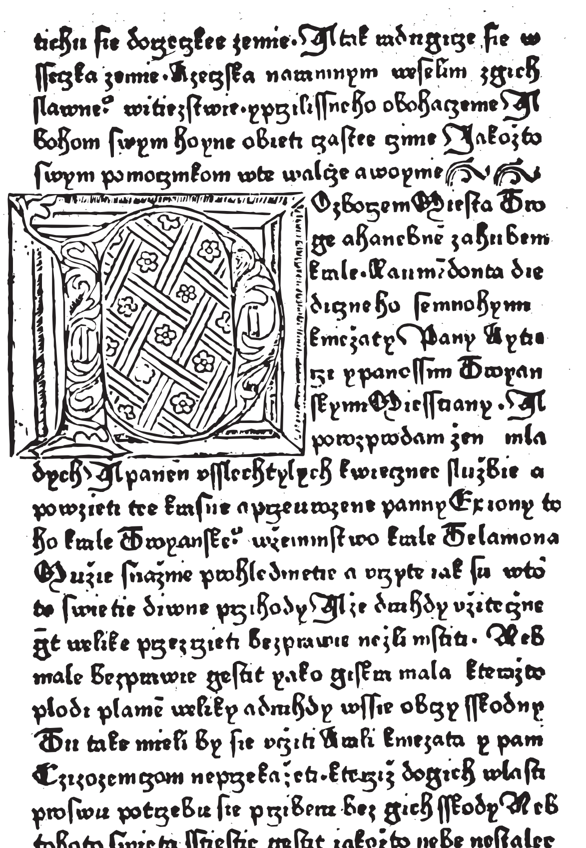

Figure 1: Page of "Kronika Trojánská"

Polska Grupa Użytkowników Systemu TEX, 1999 (http://www.GUST.org.pl)

ABCDEFGHIJK
LMNOPQRSTUVWXYZ&
1234567890
abcdefghijklmnopqrst
uvwxyz
áäàčďířšťůžfffifl ß
{.:+§„—"!?*;,}

Figure 2: Manuscript

ABCDEFGHIJ
KLMNOPQRSTUVWXYZ&
1234567890
abcdefghijklmnopqrs
tuvwxyz
[.:?áéíóúýčěňřšžďť!;,]

Figure 3: Figural

á Á ú Ú

Čárka (acute) is used to prolong the affected vowel. In "static" typefaces it usually stems from the center of the letter, in "dynamic" types of handwritten character it often "lays" on the letter – see the pictures.

ů Ů

Kroužek (ring) on "u" has the same meaning as acute, i.e, makes the vowel longer. The difference is only in spelling — ů is used inside and at the end of words (as in "růže" — rose), ú at the beginning (úhel — angle).

š Š ď Ď

Háček softens the letter: Czech "š" sounds exactly the same as Polish "sz". Note that the little thing appended to the lowercase "d" is háček, not an acute! Foreigners tend to treat it as acute and try to write it as `\'d`, which is completely wrong. The letter "t" can also be endowed with háček that has a similar property. When composed in PostScript virtual fonts, háček on these letters is made as slightly shrinked apostrophe. As you can see on the examples, háček on dynamic typefaces looks not like a reversed circumflex, but is rather "shaded". Very few fonts obey this property. Lowercase "d" with háček is also quite problematic letter, when followed by a letter that has upper long stem — they interfere in a unpleasant way:

loďka

For this reason, the háček should be as thin as possible.

CSTeX

Czech and Slovak TeX Users' Group (CSTUG) main activity is to maintain a localized Czech & Slovak distribution of TeX, that is called CSTeX. In this paper, I shall describe the platform independent parts of CSTeX — CS-fonts with Czech & Slovak letters, hyphenation patterns and localized plain and LaTeX formats. CSTeX also contains some additional programs, mainly for DOS: for example, a menu driven environment for editing and other TeX-related processing, called MNU. CS-fonts and CS-plain are in state similar to CM and TeX itself — they are frozen and "perfect" and no further changes except bug fixing are not to be made. This does not apply to CS-LaTeX, as the LaTeX itself is in development and also because it is maintained by me, therefore it is buggy. The main author and maintainer of CSTeX is Petr Olšák, who is already well known to BachoTeX auditory.

CS-fonts

The basic requirement of TeX localization is availability of suitable fonts, first of all the equivalent of TeX's basic typefaces, the Computer Modern family. For this purpose, Czech TeX contains CS-fonts, a superset of Computer Modern with accented letters added. Encoding of the lower 128 positions in CS-fonts is equivalent to Knuth's CM encoding (called `OT1` in LaTeX). The accented letters in the upper half are then encoded accordingly to ISO 8859-2. Several additional characters (as the promile sign and various quotes) are placed at some remaining positions. The complete layout of `cmr10` can be seen on figure 5.

Polska Grupa Użytkowników Systemu TeX, 1999 (http://www.GUST.org.pl)

The filename of every CS-font begins with prefix `cs`. The fonts are then called e.g., `csr10`, `csti12`, etc. The only CM fonts that don't have their CS equivalents are the mathematical ones — `cmmi`, `cmsy` and `cmex`, as there is no need for them. The METAFONT source files `csxxx.mf` all contain the same code:

```
input cscode
use_driver;
```

File `cscode.mf` defines the macro `use_driver` that determines the name of the font to be generated from the `jobname` and accomplishes that all necessary parameter files are included. Most of the code are Knuth CM files (`roman.mf`, `greeku.mf`, `punct.mf`, …) and the rest are accented and other special characters (files `csacutl.mf` for lowercase acuted letters, `cshachel.mf` for those with háček, `csadded.mf` for quotes, etc.)

The fact that most of the `mf` files are same is that on "real operating systems" they can be hardlinked into one file and save space and maintaining. On the contrary, the authors of EC fonts were not so clever and their files waste several megs of disk space. The whole described idea of the METAFONT implementation is not of Czech provenience — its authors are Bogusław Jackowski and Marek Ryćko.

Another alternative for Czech TeX users are T1-encoded EC fonts. In my opinion, they have some drawbacks compared to CS-fonts: they are much bigger and their METAFONT generation takes considerably longer time, the accents on uppercase letters are too high and cause problem in typesetting with small baselineskip. Their advantage is a better international compatibity. So, they are a good alternative to CS-fonts, are used mainly in the vicinity of Brno and the question "CS versus EC" causes occasional flamewars on the CSTUG mailing list. You can compare several characters from both fonts on figure 4.

é É č Č ť Ť ů Ů

`csr10`

é É č Č ť Ť ů Ů

`ecrm1000`

Figure 4: Comparison of CS and EC fonts

When using CS-fonts with their ISO Latin encoding, it is a good thing to have similarly encoded PostScript virtual fonts. We have them and in K. Berry's font naming convention they have a postfix `z` for common fonts and `u` for typewriter fonts (with some different positions). For example, CS-encoded Times is called `ptmr8z`.

CS-plain

CS-plain is a simple extension of Knuth's plain formats that makes use of Czech hyphenation patterns and CS-fonts. When creating the format, it first redefines the `\font` primitive in such a clever way that it inputs a `csr10` metric instead of `cmr10`, sets the `\cat/uc/lccodes` of eight-bit characters and then inputs the `plain.tex` file. Finally it loads the hyphenation patterns and dumps the format file. By default, the format behaves like original plain, but after issuing a `\chyph` it switches the hyphenation patterns, sets the `\frenchspacing` on and we can speak Czech.

Because of encoding compatibility of CM and CS-fonts (see above), no further changes must be made and low-level and plain work with CS-plain is very comfortable. When one uses an Unix-like system, where the native Czech encoding is ISO 8859-2, I don't even need any recoding system as encTeX or TCX tables, that were a big issue of the Eastern-Europe TeX world last time.

CSLaTeX

CSLaTeX is a bit larger patch that intrudes into LaTeX through its configuration interface. For example, LaTeX's hyphenation info is held in file `hyphen.ltx`, but when creating the format file, LaTeX first looks for a "configuration" file `hyphen.cfg` and it inputs its own `hyphen.ltx` only if `cfg` file doesn't exist. Needless to say, CSLaTeX provides a `hyphen.cfg` file.

Macros for loading and handling hyphenation patterns are quite complicated and much more powerful than those in the Babel system, as they give an orthogonal feeling to switching languages and font encodings. You can start the document with T1-Czech, then switch the encoding to IL2 and CSLaTeX automatically changes the `\language` for you to IL2-encoded Czech hyphenation patterns.

An "out-of-the-box" CSLaTeX format file is perfectly compatible with original LaTeX (e.g., it gives exactly the same DVI output) — the only things that are compiled into the format are declaration of ISO 8859-2 encoding (called `IL2` in LaTeX) and hyphenation patterns. To get the full Czech environment, you must

```
\usepackage{czech}
```


Polska Grupa Użytkowników Systemu TeX, 1999 (http://www.GUST.org.pl)

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x0x	Γ	Δ	Θ	Λ	Ξ	Π	Σ	Υ	Φ	Ψ	Ω	ff	fi	fl	ffi	ffl
0x1x	ı	ȷ	`	´	ˇ	˘	¯	˚	¸	ß	æ	œ	ø	Æ	Œ	Ø
0x2x	-	!	”	#	$	%	&	’	(	)	*	+	,	-	.	/
0x3x	0	1	2	3	4	5	6	7	8	9	:	;	¡	=	¿	?
0x4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0x5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	“	]	ˆ	˙
0x6x	‘	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
0x7x	p	q	r	s	t	u	v	w	x	y	z	–	—	˝	˜	¨
0x8x														‰		
0x9x									À				-	˛	«	»
0xax						Ľ				Š		Ť			Ž	
0xbx						ľ			à	š		ť			ž	
0xcx	Ŕ	Á			Ä	Ĺ			Č	É			Ě	Í		Ď
0xdx			Ň	Ó	Ô		Ö		Ř	Ů	Ú		Ü	Ý		
0xex	ŕ	á			ä	ĺ			č	é			ě	í		ď
0xfx			ň	ó	ô		ö		ř	ů	ú		ü	ý	„	“

Figure 5: Layout of CS-fonts (`csr10`)

This style file contains all the necessary definitions for Czech caption, math conventions, Czech quotes (to get a text in Czech quotes, we have the command `\uv{text}` — “uv” is a shorthand for “uvozovky” — quotes) etc.

In summary, CSLaTeX consists of the following parts:

- `cfg` files `hyphen.cfg` and `fonttext.cfg` to be compiled into the format;
- `il2enc.def` and `il2*.fd` font definition files for the ISO Latin encoded CS-fonts;
- and style files `czech.sty` and `slovak.sty` to be used by document authors.

As stated before, CSLaTeX development is not finished and it should aim to incorporation into the Babel system — it was developed separately and after improving Babel's hyphenation macros it can be united with Babel. This will mainly greatly simplify incorporation of CSTeX into thing like TeXlive and its installation into packages like teTeX — users usually first install the non-localized distribution from some CD and then patch it with the Czech stuff.

For further reading

1. Some textbook on the Czech language to be able to read the following books.
2. Petr Olšák, *Typografický systém TeX*. CSTUG, 1995, for thorough description of TeX and CSTeX written by our guru.
3. Oldřich Menhart, *Tvorba typografického písma.* SPN Praha, 1956, for a textbook on Czech typeface design.

Polska Grupa Użytkowników Systemu TeX, 1999 (http://www.GUST.org.pl)