Konverze Word Perfect $\Longrightarrow$ LaTeX

Když člověk jednou propadne TEXu (LaTEXu, $\mathcal{AMS}$-TEXu, ...), říká si, že konverze jsou zbytečné. Nejjednodušší je napsat vše rovnou v TEXu. S tímto názorem souhlasím, jenže život není žádná humanita. Často se stává, že získáme soubor, který někdo napsal nějakým textovým editorem, a my jej máme vysázet v TEXu. V tomto článku bych se chtěl podělit o své zkušenosti (a trápení) s konverzí českých textů z Word Perfectu 5.1 do LaTEXu.

K této problematice jsem se dostal tehdy, když jsem získal české texty z jistého semináře na VŠE psané ve Word Perfectu. Věděl jsem z FAQ[1], že existuje konverzní program `WP2LATEX`. Původně jej napsal v Turbo Pascalu R. C. Houtegen z TUE–Eindhoven. Proto jsou některé komentáře psány holandsky. Později jej pomocí `p2c` zkonvertoval Glenn Geers ze Sydney University do jazyka C a poté jej modifikoval Michael Covington z University of Georgia. Jak se dalo předpokládat, program neumí konvertovat české znaky. Dlouhé samohlásky byly nahrazeny správnými příkazy jako `\'e`, neboť se vyskytují např. ve francouzštině. Většina českých znaků se však zkonvertovala na `\P`, což není moc užitečné. Než ale popíšu, jak jsem problém vyřešil, budu pokračovat v líčení dalších potíží.

Z dokumentu z VŠE jsem si tedy zjistil kódování češtiny ve Word Perfectu 5.1. Dlužno poznamenat, že to nebyl ani kód Kamenických ani Latin2 ani KOI8ČS. Původně jsem chtěl svoji verzi dát k dispozici na nějaký server, ale text z VŠE neobsahoval všechna česká písmena a konverzní funkce tedy nebyla kompletní. Proto jsem od svého záměru upustil.

Po krátkém čase mi volal známý, že v jistém pražském nakladatelství mu tisknou cosi napsané českým Word Perfectem 5.1[2] a bohužel se jim porouchala tiskárna. Hledá tedy někoho, kdo by mu to nějak vytiskl. Proto jsem navštívil jinou známou, která pro jiné nakladatelství psala texty českým Word Perfectem 5.1. Napsal jsem si u ní na disketu všechna česká písmenka, abych mohl dokončit konverzní program. K mému velkému zděšení měl její český Word Perfect 5.1 úplně jiné kódování češtiny než Word Perfect 5.1 z VŠE. Nyní vás asi nepřekvapí, že známý, pro kterého jsem si konverzní program připravoval, měl ve své verzi Word Perfectu 5.1 kódování zcela odlišné...

Ideu obecného konverzního programu jsem tedy vzdal. Z programu `WP2LATEX` jsem vytvořil pomocný nástroj, s nímž lze konverzi provést, avšak definici kódování je nutno upravit podle konkrétního dokumentu.

Teď už se konečně dostaneme k výkladu, jak je vše uděláno. Word Perfect má řadu znakových sad (např. znaková sada 8 obsahuje řeckou abecedu). Konverzní program tedy musí rozeznat nejenom kód znaku, ale i sadu. Pro tento účel si `WP2LATEX` vytváří tabulku. Ta je definována ve funkci:

```
Local␣Void␣Ext_chr_init(LINK)
```

Funkce obsahuje řadu příkazů. Pro ilustraci uvedu následující:

```
LINK->char_set[0x1c]␣=␣0x4;
LINK->char_code[0x1c]␣=␣0xb;
strcpy(LINK->ext_lat[0x1c],␣"\\pounds␣");
```

[1] Máte-li přístup na INTERNET, víte, že zkratka FAQ znamená „Frequently Asked Questions“ a příslušný dokument obsahuje nejčastěji kladené otázky týkající se TEXu a samozřejmě též odpovědi.

[2] Všimněte si, že stále mluvím o „stejné“ verzi Word Perfectu!

Hodnota 0x1c nemá žádný hlubší význam, je to pouze index v konverzní tabulce. Uvedené příkazy specifikují, že znak ze sady 0x4 s kódem 0xb se nahradí řetězcem `\\pounds␣`, tj. vytiskne se jako £. Pokud se nějaký znak v tabulce nenajde, v LaTeXovém souboru se objeví `\P`. Ošetření tohoto případu se vyskytuje ve funkci

```
Local␣Void␣Character(LINK)
```

v následujících příkazech:

```
if␣(found)
␣␣␣strcpy(ch,␣LINK->ext_lat[j␣-␣0x80]);
else
␣␣␣strcpy(ch,"\\P");
```

V programu jsem udělal dvě úpravy. Nejprve jsem zrušil všechny příkazy ve funkci `Ext_chr_init`. Abych dostal něco užitečnějšího než strohé `\P`, nahradil jsem poslední příkaz výše uvedeného příkazu `if` příkazem:

```
␣␣␣sprintf(ch,␣"\\P{%x/%x}",␣(unsigned)char_set,␣(unsigned)char_code);
```

Tím získám ve výpisu jak číslo znakové sady, tak kód nepoznaného znaku. Kdyby se v textu vyskytoval £ a nebyl v tabulce, měl bych ve výstupním souboru `\P{0x4/0xb}`. Po dohodě s autorem textu nebo často i intuicí lze kódování rozluštit a doplnit do funkce `Ext_chr_init`. Do konverzní tabulky lze zadat až 256 znaků.

`WP2LATEX` má svůj vlastní „style option“ pro tisk LaTeXem. Nezkoušel jsem konvertovat matematiku ani tabulky, takže nevím, jak se v těchto případech konverzní program zachová. Při konverzi hladkých textů doplňuje určitá svá makra, ale naštěstí je jich tak málo, že je lze v libovolném editoru snadno ručně odstranit a dokument pak lze snadno vytisknout i v plain TeXu.

Zdeněk Wagner
wagner@icpf.cas.cz