Zpravodaj Československého sdružení uživatelů TeXu

Jaroslav Fojtík

Konverze dokumentů z WordPerfectu do LaTeXu – WP2LaTeX

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 8 (1998), No. 2, 97–104

Persistent URL: http://dml.cz/dmlcz/149816

kud máte pěkně počeštěné písmo pouze v Corku, nevadí, mám awk skript, který převede virtuální fonty do CS-kódování.) Pokud najdete na stránkách nějaké chyby, dejte mi, prosím, vědět. Pokud dokážete zveřejněná písma klasifikovat podle ON 88 1101, tj. podle třídění Jana Solpery, prosím, pošlete mi tyto klasifikace.

Adresa Fontanasie je `http://www.econ.muni.cz/~qasar/fontanasia/`.

Přeji příjemné TEXování.

Michal Kvasnička
`qasar@econ.muni.cz`

Konverze dokumentů z WordPerfectu do LaTeXu – WP2LaTeX

Jaroslav Fojtík

Tento článek se zabývá konverzí textových dokumentů z textového procesoru Word Perfect v 5.x do typografického systému LaTeX. Autor v něm popisuje vlastní vylepšenou verzi původního programu, která umí konvertovat akcentované znaky (tedy upravená pro češtinu) a navíc přidává konverzi dalších speciálních rysů, např. tabulek a matematických formulí.

1. Úvod

WP51 byl před několika lety nejúspěšnější textový editor. Byl znám hlavně v USA, ale i v Čechách byl hojně používaný. Například všechny knížky vydavatelství GRADA v něm byly vytvářeny. Takto jsem se s ním seznámil. Poté jsem byl donucen začít psát v LaTeXu. Proto jsem přemýšlel, jestli by bylo možno původní texty ve WP nějakým způsobem převést do LaTeXu. Nalezl jsem konvertor na internetu a pokusil se provést převod textu. Protože výsledky konverze nebyly pro mne uspokojivé, tak jsem se jej rozhodl upravit. Navíc jsem si přál poznat vnitřní strukturu souborů WP.

2. Historie programu

Znalci archivu CTAN mohou být překvapeni. Na co další nový konvertor, když už jsou v archívu dva dostupné? Proč nevyužít rovnou některého z nich? Pokusím se tedy odpovědět.

První konverzní program, jehož autorem je R. C. Houtepen, je poměrně starý a umí konvertovat velmi málo rysů WP. Hlavním a největším problémem původního konverzního programu při konverzi dokumentů v češtině/slovenštině (a jiných jazycích používajících písmena s akcenty) je, že tyto znaky prostě ignoruje. Autora nelze kontaktovat, a proto od tohoto programu bylo odvozeno několik dalších mutací. Za základ popisovaného programu byl vzat právě tento kód. Avšak byl téměř celý přepsán.

Druhý dostupný program vznikl převedením původního programu `WP2LATEX` z PASCALU do jazyka C++ a přidáním možnosti konvertovat rovnice. Na úpravách se podíleli autoři Glenn Geers, Michael Covington, Claudio Porfiri a Dirk Lellinger. Ani tento program však neřeší problém češtiny korektně. Navíc má jednu odpornou vlastnost, že při konverzi složitějších rovnic padá. Autoři to nazývají Force Lazy Bracketing (Vnucování Líného Závorkování). Po komunikaci s jedním[1] z autorů jsem se dozvěděl, že s tímto produktem již nechtějí mít nic společného.

Samozřejmě, že mohou existovat konverzní programy z WP do LaTeXu od jiných autorů. Za zmínku stojí program `wp2x`, který je velmi obecný a umožňuje konverzi z WP do HTML, LaTeX a jiných systémů. Je sice velmi obecný, ale jeho konverzní možnosti se omezují jen na nejjednodušší příkazy pro formátování dokumentů. Je volně k dispozici na `sunsite.unc.edu` včetně zdrojového kódu v jazyce C.

Dalším programem podobného zaměření je `texperfect` of Johna Forkoshe. Vzhledem k dosažené kvalitě konverze se k němu raději vyjadřovat nebudu.

Když jsem se dozvěděl o vzniku tolika derivátů původního programu, rozhodl jsem se vytáhnout ze šuplíku svoji upravenou verzi, dodělat do ní všechny nové rysy, které jsem v různých mutacích nalezl, poslat ji do archívu CTAN a nabídnout ji v tomto časopise potenciálním zájemcům.

Ať se na mne autoři ostatních konvertorů WP nezlobí, ale zde popisovaný program je s nimi nesrovnatelný. Pokouší se o řádově vyšší způsob konverze se kterým jsem se zatím nikde nesetkal. Navíc jako jediný je schopen (s jistými omezeními) zpracovat dokumenty Word Perfectu verze 4.x a verze 6.x (Tato vlastnost je dostupná od verze WP2LaTeX 2.43). Domnívám se, že by bylo lepší spojit úsilí a dodělat zbylé zatím nekonvertované vlastnosti WP (např. čínské znaky, speciální symboly, čtverečky) než vyvíjet další vlastní jednoduchý konvertor.

[1] Nepsaným konsensem je, že se o program vždy stará ten, kdo provedl modifikaci jako poslední.

3. WP2LaTeX v akci

Popisovaný program je napsán v jazyce Pascal a je nabízen i se zdrojovými texty. Dále již je plně funkční verze v C++. Z důvodu jistých omezení Pascalu již tato verze nebude dále rozvíjena. V balíku `wp2latex.zip` je konvertor přeložen pro použití v operačních systémech MSDOS, Linux, OS/2 Warp 3.0 a Windows 3.11.

Syntaxe programu:

```
WP2LATEX [WP-filename.WP [TeX-filename.TEX]] [/cp895 | /cp852]
  [/texchars] [/charset1 | /charsetCZ] [/optimizesection]
  [/?] [/silent]}
```

Základní spuštění programu pro první experimenty je velmi jednoduché. Stačí pouze odeslat z příkazové řádky jméno programu bez argumentů. Ten si sám požádá o zadání vstupního souboru a následně nabídne soubor výstupní. Stačí odklepnout klávesu `ENTER` a po krátké době je již vytvořen výstupní soubor pro LaTeX. Pro první pokus můžete použít například soubor `test.wp`. Výstupní soubor konvertoru již může být rovnou překládán LaTeXem. Pro překlad bude vyžadován ještě soubor `wp2latex.sty`, který je součástí základního balíku.

Neočekávejte lepší vzhled dokumentu vzniklého pouze konverzí z WP do LaTeXu! Nejspíše budete muset provést ruční editaci dokumentu. Důvodem této skutečnosti je, že LaTeX obsahuje hlavně příkazy pro nastavování struktury dokumentu a slovní procesor WP obsahuje spíše příkazy pro změnu vzhledu textu. Někdy bude též vhodné editovat původní WP dokument a odstranit z něj zbytečné řídící znaky. To jsou ty znaky, které se v tištěné podobě dokumentu neobjeví, ale jsou stále v souboru (např [tab] na konci řádky). Nebudou-li výsledky konverze uspokojivé na první pokus, je možné zkusit spustit konverzní program s jinými parametry.

Za zmínku stojí, že vlastní konverze je prováděna ve dvou průchodech. V prvním průchodu je prováděna detekce prostředí a je konvertována většina rysů WP. Ve druhém průchodu je prováděna optimalizace textu, doplňován úvod a zakončení sekcí a dokumentu.

Pro vaši představu nyní uvedu seznam rysů, které jsou programem WP2LaTeX akceptovány. Podotýkám, že ne všechny vlastnosti vyjmenovaných řídících kódů jsou konvertovány na 100 %.

- Posunutí (doleva, doprava, nahoru, dolu) `[AdvDn] [AdvUp] [AdvLft] [AdvRtg]`
- Křížové odkazy
- Rovnice
- Rozšířené znaky (cizí jazyky/akcenty, matematické symboly, řecká abeceda - ne čárové kresby)
- Zarovnání textu napravo `[Fls Rgt]`

- Poznámky pod čarou `[Footnote]`
- Poznámky na konci textu `[Endnote]`
- Umístění textu :
  - vlevo `[Just Left]`
  - napravo `[Just Right]`
  - centrované `[Just Center]`
  - plné `[Just Full]`
- Vynucený konec stránky (Hard Page) `[HPg]`
- Vynucený konec řádky (Hard Return) `[HRt]`
- Záhlaví a Zápatí `[Header]` `[Footer]`
- Odsazení :
  - vlevo `[->Indent]`
  - vlevo a vpravo `[->Indent<-]`
- Obrysové písmo (Outlining)
- Přepis (Overstrike) `[Ovrstk]`
- Umístění čísel stránek
- Tabelátory : `[Tab]`
  - nastavení
  - levé tabelátory
  - centrované tabelátory
  - pravé tabelátory
- Tabulky
- Následující tvary a velikosti písem :
  - Zvláště velké (Extra large) `[Ext Large]`
  - Velmi velké (Very large) `[Vry Large]`
  - Velké `[Large]`
  - Malé `[Small]`
  - Drobné písmo `[Fine]`
  - Horní index `[Supscript]`
  - Dolní index `[Subscript]`
  - Kurzíva `[Italic]`
  - Kapitálky (Small caps) `[Sm Cap]`
  - Tučné `[Bold]`
  - Podtržené (Underlined) `[Und]`
  - Dvojitě podtržené (Double-underlined) `[Dbl Und]`

4. Rozbor dílčích aspektů konverze

4.1. Konverze češtiny

Tomuto tématu se budu nyní věnovat o něco podrobněji. Problémy, které zde bylo potřeba řešit jsou celkem dva. Je používáno u nás velké množství způsobů

kódování češtiny. Tento problém vlastně TEX již řeší sám o sobě pomocí svého způsobu zápisu akcentovaných znaků. Proto program WP2LATEX provádí implicitně konverzi právě do sedmibitového TEXu. Na požádání umí navíc výstupní kód vytvořit v osmibitovém kódování. Znaky s akcenty, které se v tomto kódování nevyskytují, jsou samozřejmě ponechány v notaci TEXu. Pro přepnutí do tohoto kódování lze použít následujících přepínačů.

`/cp895` Zvolí výstupní kódování Kamenických.
`/cp852` Zvolí výstupní kódování ISO8 (Latin 2).

Dalším problémem, který vzniká při převodu akcentovaných znaků, je několik interních reprezentací češtiny. Pro řešení tohoto problému jsem navrhl přepínače `/charset1` a `/charsetCZ` . Předpokládám, že nejčastější je znaková sada charset 1. Pokud budete mít jinou znakovou sadu[2], zkuste mi poslat soubor `charactr.wp`.

Pro přehled o konverzních schopnostech programu doporučuji překlad souboru `charactr.wp`, který je dodán spolu s programem `wp2latex.exe`. Pak můžete rovnou na obrazovce vidět, které znaky jsou překládány, a z přeloženého souboru `charactr.tex` je možno vypátrat jak.

4.2. Matematické formule

Zápis matematických formulí ve WP je velmi podobný zápisu formulí v LATEXu. Proto, jak jsem se již zmínil, předchozí verze umožňovala provádět konverzi formulí. Použitá metoda se snažila detekovat celý výraz a podle toho vytvořit zápis v LATEXu. Podle mě byla použitá technika příliš složitá. Dokonce tak složitá, že ji ani sám autor nedokázal dovést do bezchybného stavu.

Proto ve své verzi konvertoru používám jinou (vlastní) metodu založenou na kontextové gramatice. Zjednodušeně by se dalo říci, že všechny symboly Word Perfectu představují neterminální symboly. Příkazy LATEXu jsou považovány za terminální symboly. V programu je definováno několik desítek přepisovacích pravidel, které jsou aplikovány na množinu symbolů do té doby, dokud se v množině vyskytují neterminální symboly.

Domnívám se, že není potřeba zatěžovat čtenáře dalšími informacemi tohoto rázu. Poznamenávám, že se mi zatím nepodařilo ve WP napsat takový výraz[3], který by byl překonvertován chybně. To však neznamená, že se takový výraz nepodaří najít čtenáři tohoto článku.

4.3. Tabulky

Podobně, jako LATEX umí i Word Perfect vytvářet tabulky. Tabulka ve WP vypadá asi následujícím způsobem:

[2] Lokalizátoři a tvůrci Word Perfectu, styďte se za takový zmatek!
[3] Možnosti zápisu formulí ve WP jsou o dost chudší než LATEXu.

```
[Tbl Def:I;3,5.31cm,5.31cm,5.31cm]
[Row] [Cell] Left Centered  [Cell] Center [Cell] Right Centered
[Row] [Cell] 111 [Cell] 222 [Cell] 333 Centered
[Tbl Off]
```

A její ekvivalent po konverzi je:

Left centered	Center	Right centered
111	222	333

Při konverzi tabulek nastává pouze jediný problém. WP má mnohem obecnější možnosti zápisu tabulek. Umožňuje některé čáry vynechávat, měnit pozici textu v jednotlivých buňkách, některé buňky stínovat atd. Všechny tyto rysy zůstávají nekonvertovány.

5. Speciality

/optimizesection
Program se snaží odhadovat z příkazů WP příkazy pro formátování textu. Detekce kapitol a oddílů je odvozována ze symbolů typu `[Mark: ToC ToCLevel]`. Pro definici oddílu ve WP je však potřeba udělat mnohem více. Většinou je navíc změněn typ písma, jeho velikost atd. Právě tento 'nepořádek' je po zapnutí přepínače `/optimizesection` vyhazován.

/texchars
Normálně jsou znaky typu `$, {` atd. konvertovány na `\$`, `\{`. Zapnutí uvedeného přepínače tuto volbu potlačí. Tím je umožněno používat WP jako textový editor pro LaTeX.

Editace dokumentů LaTeXu ve WP, které mohou obsahovat azbuku
Při použití parametru `/texchars` lze navíc přidávat i řídící příkazy TeXu a LaTeXu do zdrojového dokumentu WP a tyto příkazy nebudou při konverzi poškozeny. Tímto by se dalo psát rusky mnohem komfortněji než v T602 a používáním programu AZBTOTEX. Tato možnost vychází ze schopnosti konvertovat azbuku do LaTeXu. Při vhodném nastavení WP lze zobrazit současně jak azbuku tak latinku na obrazovce. Existuje způsob, jak zobrazit na textové obrazovce současně 512 různých znaků a WP jej podporuje.

Konverze dokumentů z WORDu do LaTeXu
Byl vyzkoušen způsob konverze dokumentů z WORDu[4] do LaTeXu. Ve WORDu je potřeba zapnout výstupní filtr pro WP 5.1 a uložit dokument v tomto formátu.

[4]Tento odstavec jsem do textu zařadil zejména z důvodu, že jsem byl na tuto věc již několikrát tázán.

Další postup při konverzi je již totožný s konverzí dokumentů napsaných ve WP. Tento postup byl skutečně ověřen avšak nečekejte od takto konvertovaných dokumentů zázraky. Při konverzi z WORDu do WP dochází ke ztrátě některých informací, které již nelze automaticky obnovit.

Konvertor musel být speciálně upraven pro tento způsob konverze. Word totiž na svém výstupu vytváří množství formátovacích příkazů (smetí), které nemají v dokumentu co dělat a zhoršují jeho přenositelnost.

Konverze poškozených dokumentů
Další zajímavou vlastností je možnost konvertovat poškozené dokumenty, které již nelze načíst ani do WP. Ten se při jejich načítání většinou zhroutí[5]. Pro konverzi poškozených dokumentů je potřeba zapnout přepínač `-safe-mode`. Zvýšená kontrola nepřipustí poškozené objekty ke konverzi. Zejména od této volby si neslibujte zázraky, ale někdy stojí za pokus zachránit co se dá.

Konverze dokumentů zapsaných ve WP 6 a WP 4
Tato možnost je zatím pouze experimentální. Množství konvertovaných rysů je dosud malé. Pokud je mi známo, ještě se o to nikdo nepokusil (mám na mysli hlavně WP 6).

6. Diskuse o kvalitě konverze

Když jsem se rozhodl upravit konvertor, v manuálu jsem se dočetl podivné věci. Prý je po konverzi potřeba vytvořit několik filtrů pro program SED a navíc několik programů v C. Pak je prý potřeba vytvořit si vlastní styl a již bude konvertovaný dokument vypadat skvěle.

Pokusil jsem se program upravit, aby negeneroval na svém výstupu nekorektní příkazy LaTeXu. Pokud je mi známo, žádný další autor konvertoru z WP do LaTeXu se o tuto věc ani nepokusil. Podařilo se mi provést přes 40 dílčích zlepšení, která se promítají do kvality dokumentu po konverzi. Ale dokonalosti se mi zatím dosáhnout nepodařilo.
Některé důvody jsou:

1. Prováděná konverze je pouze dílčí a některé rysy stále zůstávají ignorovány.
2. V dokumentech Word Perfektu se nacházejí nechtěné zapomenuté objekty (jako např. `[Tab]`, `[Tab Set]`, `[indent]` namísto `[Tab]` atd.). Proto doporučuji používat `[Tab]` co možná nejméně.
3. Ve Word Perfektu je povoleno dělat nečisté věci, které jsou v LaTeXu zakázány, například `\section{Some sect\tableofcontents ion}`. Jak jsem

[5] Toto jsem si prakticky ověřil po uložení své diplomové práce na vadnou disketu.

se již zmínil, těžiště všech optimalizací leží právě v odstraňování a nápravě těchto zjevů.

7. Implementace programu

Současná verze `wp2latex-2.48` je zapsána v C++. Je přeložitelná překladači Borland C++ 3.11, Borland C++ OS/2, GCC DJGPP, GCC Linux a HP UX (posledním překladačem s určitými problémy).

Celý program podporuje začlenění knihovny `gettext`, která umožňuje překládat všechna výstupní hlášení za chodu programu. V současné době mám k dispozici pouze hlášení v českém jazyce (což je pro čtenáře českého časopisu jistě dostatečné). Způsob jakým knihovna `gettext` překládá za chodu texty nepatří do tohoto článku. V případě zájmu by bylo možno o této problematice napsat jiný článek.

Pro operační systémy typu UNIX je povolen zápis argumentů začínající jak lomítkem, tak i pomlčkou.

Vzhledem k rozšíření Pascalu v jiných OS než je DOS a vzájemné nekompatibilitě překladačů Pascalu jsem od verze v Pascalu upustil. Hlavním důvodem byla nemožnost pracovat efektivně s řetězci delšími než 256 znaků.

8. Závěr

Programem WP2LaTeX dávám uživatelům nástroj pro konverzi dokumentů. Nabízený program je schopen konvertovat i dokumenty o rozsahu několika megabajtů. Tímto vám dávám možnost převést stávající dokumenty z textového procesoru WP do LaTeXu a tím je vymanit z komerčního prostředí plného pěkných obrázků a vzájemných nekompatibilit. Poznamenávám, že velké množství programů má výstupní filtr pro WP5.x, a takto by bylo možno konvertovat některé dokumenty do LaTeXu dvoustupňově.

Přeji všem uživatelům WP2LaTeXu, aby jim program dobře sloužil. Vývoj uvedeného programu nepovažuji za ukončený, a proto vítám jakékoliv další připomínky včetně spolupráce na jeho dalším vývoji. V případě nejasností nebo s připomínkami se můžete obrátit na moji Emailovou adresu `fojtik@vision.felk.cvut.cz` nebo `fojtik@cmp.felk.cvut.cz`. Nejnovější verze programu je k dispozici na mé *www* stránce `http://cmp.felk.cvut.cz/~fojtik/`. Popisovaný program je též možno získat z archívu CTAN (`CTAN:support/wp2latex`).

Jaroslav Fojtík
fojtik@vision.felk.cvut.cz