Zpravodaj Československého sdružení uživatelů TeXu

Ladislav Lhotka
České dělení pro TeX

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 1 (1991), No. 4, 8–9

Persistent URL: http://dml.cz/dmlcz/148814

České dělení pro TEX

Motto: "... patterns are supposed to be prepared by experts who are paid well for their expertise."

(D. Knuth, The TEXbook)

Začátkem měsíce listopadu jsem dokončil některé podstatnější úpravy souboru `cshyphen.tex` obsahujícího vzory pro české dělení. Děkuji tímto Karlu Horákovi, který mi předal poměrně rozsáhlý soubor se slovy, která předchozí verze dělila špatně nebo neúplně. Nová verze vzorů je k dispozici v „ústředí" CSTUGu (u Olina Ulrycha), případně je mohu přímo zaslat elektronickou poštou.

Při této příležitosti jsem se znovu rozpomněl na idylické doby vzniku původní amatérské české lokalizace TEXu. Moje první varianta vypadala tak, že jsem příslušný počet znaků ze začátku ASCII tabulky deklaroval jako aktivní (například písmeno *š* mělo kód `"08`) a jejich definice odpovídala standardním plovoucím akcentům typu `\v s`. Na prvních 32 míst se vešla všechna malá písmena s akcenty a z velkých písmen ta, která se mohou vyskytnout v běžném textu[1]—ostatní velká písmena, například v nadpisech, se realizovala pomocí maker `\v` a `\'`. Pro vkládání zdrojového textu na PC a jeho případné tištění na tiskárně ovšem toto rozložení nebylo vhodné, protože některé z inkriminovaných kódů mají řídicí funkce, třeba kód `"1B` (ESC) zapíná povelový režim většiny tiskáren. Zde byla pomoc snadná: Z hlediska počítače se všechny kódy posunuly 0 128 pozic výše, kde nepůsobily problémy. Sedmibitový TEX pak nejvyšší bit uřízl a všechno bylo v pořádku. Jediným problémem bylo dělení, protože před verzí 3.0 nedělil TEX slova obsahující aktivní znaky nebo makra.

Někdy na podzim roku 1988 jsme se domluvili s Petrem Novákem a začali uvažovat o „profesionalizaci" češtiny (a slovenštiny) v TEXu (oba jsme tehdy byli aspiranty a měli proto na takové věci čas). Přes značné odhodlání jsem nakonec neudržel své kódování češtiny v PC proti standardu Kamenických, čímž ovšem vyvstala nutnost filtrovat vstupní soubory (tedy to, co emTEX zařizuje sám pomocí `tcp` tabulek). Vyjasnili jsme si také, že slušné české dělení je možné jen tehdy, vyrobíme-li TEXovské fonty obsahující přímo akcentovaná písmena, a dále rozsáhlejší soubor vzorů pro české dělení. Vzhledem k tomu, že jsem měl k dispozici METAFONTbook, dohodli jsme se, že já připravím fonty a Petr se bude věnovat dělení.

Vše nakonec dopadlo zcela jinak. Petr si okopíroval METAFONTbook a, propadnuv jeho nespornému kouzlu, začal experimentovat s písmenky. Takto postupně vznikly jeho CS fonty, které byly první (a jsou zřejmě dosud jedinou) kvalitní československou mutací rodiny *Computer Modern*. No a na mne tedy zbylo dělení.

Půjčil jsem si v knihovně prastarou knihu Jiřího Hallera *Jak dělit slova* (SPN 1956) a začal studovat. První kroky byly snadné. Základní vzory dělení v češtině i slovenštině jsou typu, *samohláska–souhláska, samohláska a samohláska, souhláska–souhláska, samohláska*. Napsal jsem proto program, který tyto kombinace generoval. Přímo v programu a pak i ručně jsem eliminoval nemožné skupiny hlásek typu *ůxů* (doufám, že v tomto článku nebude nutno tuto skupinu dělit). Nejobtížnější částí bylo zpracování předpon, kde se možnosti dělení zpravidla rozsáhle větví (*pře–, před–, přede–*). Musel jsem projít celý (nepříliš rozsáhlý) slovník v Hallerově knize a uvážit

[1] nikoliv tedy např. Ě

všechny možné varianty a výjimky. Podobné to bylo s příponami. Nakonec přišly na řadu cizí předpony a přípony a izolované výjimky. Přibližně po dvou týdnech usilovné práce byla první verze hotova.

Počáteční fáze testování a úprav ukázala některé důležité věci:

1. Ve spojení s TEXem umožňuje tato tabulka dělení docela uspokojivě sazbu českého textu bez nutnosti příliš častých zásahů
2. Protože je soubor uspořádán do logických celků, je poměrně snadné doplňovat postupně další vzory a uvážit všechny možné vedlejší efekty.
3. Největší problémy jsou ve slovech složených anebo obsahujících více předpon (například *nejneobvyklejší*). Tato negativní vlastnost v podstatě přetrvává i ve všech dalších verzích. Naštěstí nemáme v našich jazycích takovou frekvenci složenin jako třeba němčina (*Donaudampfschiffsfahrtkapitänkajütentürschlo*).

I přes původní záměr jsem nakonec rezignoval na současné zpracování slovenské částí vzorů. Správně jsem předpokládal, že se tohoto úkolu zhostí dříve či později někdo ze Slovenska (učinila tak Janka Chlebíková).

S výsledkem jsem byl s ohledem na vložené úsilí velice spokojen. Ukázalo se, že v mnoha případech automatické dělení překonalo svého tvůrce, když jsem ke svému překvapení musel po prověření ve slovníku opravit své mylné představy. Postupně se sice objevovala některá více či méně fatální opomenutí (kupříkladu *Česko-slovensko*—slovenští přátelé, neříkejte to na mě vašim pomlčkovým mamelukům), která jsem byl, většinou bez větší námahy, schopen průběžně doplňovat.

Větší problémy mi paradoxně přinesl až TEX 3.0, který mimo jiné odstranil tvrdé pravidlo předchozích verzí, že totiž ze slova lze oddělit nejméně dvě písmena na začátku a tři na konci. S tímto omezením jsem mlčky počítal při přípravě vzorů. Po nastavení parametrů `\lefthyphenmin` a `\righthyphenmin` na místním zvyklostem odpovídající hodnoty 1 a 2 jsem s hrůzou zjistil, že se dělí třeba *h-růza*. Proto bylo třeba podniknout rozsáhlejší revizi vzorů.

V poslední verzi jsem zejména doplnil některé časté dvojice předpon (nepřeapod.), další cizí předpony, složené číslovky a jiná důležitá složená slova. Karel Horák mi doufám odpustí, že případy jako *SNTL*, *JČMF* či *Tartaglia* si bude muset nadále ošetřovat sám.

Je zřejmé, že ani poslední verze není dokonalá. Pokud tedy naleznete případy chybného dělení či nedělení, podejte mi prosím zprávu. Rychlost odezvy ovšem nemohu garantovat, nejsem už totiž aspirant. Výhledově by bylo pravděpodobně užitečné připravit w nejúplnější soubor vzorů pomocí speciálního programu PATGEN na základě nějakého velkého slovníku. Petr Novák mezitím něco takového udělal, svůj výsledek však zatím zřejmě neposkytuje volně TEXové komunitě. Momentálně se tedy musíme smířit s občasnou ruční opravou případů, kde můj skromný (zato ale *public domain*) příspěvek selhává. Není vyloučeno, že jím nepohrdnou ani Lidové noviny—myslim si, že by to bylo každopádně lepší než jejich současné dělení.

(*Ladislav Lhotka*)
e-mail: `lhotka@csearn`