

# Zpravodaj Československého sdružení uživatelů TeXu

---

Martin Budaj

Divide et impera: program findhyph

*Zpravodaj Československého sdružení uživatelů TeXu*, Vol. 20 (2010), No. 1-2, 2-5

Persistent URL: <http://dml.cz/dmlcz/149986>

## Terms of use:

© Československé sdružení uživatelů TeXu, 2010

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* <http://dml.cz>

## Abstrakt

Tento článok popisuje jednoduchý program `findhyph`, ktorý v ľubovoľnom dokumente spracovanom v  $\text{T}_{\text{E}}\text{X}$ u nájde všetky slová, ktoré boli rozdelené pri zalamovaní odsekov. Program je dostupný na `CTAN.ORG`.

**Kľúčové slová:**  $\text{T}_{\text{E}}\text{X}$ , Perl, kontrola rozdelených slov, nástroj `findhyph`.

doi: 10.5300/2010-1-2/2

## Divide: použitie programu

Zoznam rozdelených slov je užitočnou pomôckou pri kontrole dokumentov, na sadzbe ktorých nám obzvlášť záleží; najmä ak sú napísané v jazyku, pre ktorý nie sú dosiaľ vytvorené spoľahlivé vzory delenia. Program `findhyph` vytvorí takýto zoznam na základe informácií v log-súbore vytvorenom  $\text{T}_{\text{E}}\text{X}$ om pri sadzbe dokumentu.

Okrem rozdelených slov nájde program aj jednohláskové predložky a spojky ponechané v rozpore so základnými pravidlami sadzby podľa ON 88 2503 na konci riadku [1]. Je to častý jav, pokiaľ na ich pripájanie k nasledujúcim slovám nepoužívame makro `~` (vlnovka).

Tento článok simuluje použitie nedokonalých vzorov delenia pomocou použitia českých vzorov pre slovenský text. Pre názornosť sú všetky miesta riadkového zlomu, ktoré sú zaujímavé pre program, vyznačené šípkou na okraji textu.

Použitie programu je ilustrované v nasledovných odsekoch.

Predpokladajme, že pracujete so súborom `subor.tex`. Na jeho začiatku je potrebné uviesť príkaz `\tracingparagraphs=1`. Potom ho spracujte  $\text{T}_{\text{E}}\text{X}$ om alebo `pdfTEX`om bežným spôsobom. Na použítom formáte nezáleží – program bol úspešne testovaný s `plain TEX`om, `LATEX`om aj `ConTEX`tom.  $\text{T}_{\text{E}}\text{X}$  vytvorí `subor.log` s informáciami o možnostiach riadkového zlomu [4, 6].

Teraz treba spustiť program `findhyph` s argumentom `subor.log`. Program z údajov log-súboru zrekonštruuje rozhodovanie  $\text{T}_{\text{E}}\text{X}$ u pri zalamovaní riadkov a nájde všetky rozdelené slová. Tieto sú uložené do súboru `subor.hyph`. Rozdelené slová je možné zobrazíť aj v kontexte spolu s predchádzajúcim a nasledujúcim slovom, pokiaľ je program spustený s parametrom `-c`.

Ak je program spustený s parametrom `-p`, sú nájdené aj jednohláskové predložky a spojky zabudnuté na konci riadka; uložené sú do súboru `subor.prep`. Program štandardne uvádza všetky predložky a spojky okrem malého písmena *a*. Parameter programu `-l` umožňuje užívateľsky predefinovať zoznam hľadaných predložiek a spojiek.

Príklad použitia programu na kontrolu tohto článku:

```
tex clanok.tex
findhyph -cpl=kKsSvVzZoOuUiIaA clanok.log
```

Program vytvorí súbory `clanok.hyph` a `clanok.prep`. Číslo v hranatých zátvorkách pod textom označuje číslo strany, ktorú  $\text{T}_{\text{E}}\text{X}$  zapísal do DVI alebo PDF súboru po zalomení daného textu. Keďže  $\text{T}_{\text{E}}\text{X}$  pred zápisom strany do DVI alebo PDF súboru môže do riadkov zalomiť viac odsekov, ako sa na stranu nakoniec vojde, nemusí byť číslovanie strán v súboroch vytvorených programom `findhyph` vždy zhodné s číslovaním strán vo vysádzanom dokumente. V prípade výskytu napríklad plávajúcich obrázkov sa zalomený text môže objaviť vo vysádzanom dokumente až o niekoľko strán ďalej.

Poznámky pod čiarou spôsobia iné poradie textov v súboroch `clanok.hyph` a `clanok.prep` oproti vysádzanému dokumentu, keďže v log-súbore je uvedená najprv informácia o zalomení poznámky pod čiarou a až následne informácia o zalomení odseku, v ktorom sa poznámka vyskytuje. Obdobne sú rozdelené slová z popisu k plávajúcemu obrázku uvedené na mieste vloženia obrázku v zdrojovom dokumente, nie na mieste finálneho umiestnenia obrázku po vysádzaní.

clanok.hyph

```
ľubovoľnom do-kumente spracovanom
a ná-jde všetky
jednohláskové pred-ložky a
[2]

interpunkčné znami-enka, zátvorky
[4]

systéme Win-dows) skopírovať
Slovak Type-setting Rules.]
z ad-resára: http://ctan.org/texarchive/support/findhyph/[]
[5]
```

clanok.prep

```
Program z údajov
[2]

informácia o zalomení
[3]

zobrazenie v textovom
popísané v knihách
[4]
```

Súbor s rozdelenými slovami obsahuje pre informáciu aj interpunkčné znamienka, zátvorky a ostatné znaky bezprostredne predchádzajúce alebo nasledujúce slová uvedené v súbore.

Text nemusí byť zobrazený úplne, pokiaľ je príliš zložitý na zobrazenie v textovom log-súbore – napríklad ak je jeho časť uzavretá v primitíve `\hbox{}`, alebo ak obsahuje značku `\mark{}`. Zložitá časť sa zobrazí ako prázdne hranaté zátvorky, napríklad makro `\TeX` sa zobrazí ako `T[]X`.

Prechod do matematického módu je vyznačený symbolom `$`. Zmeny typu písma nie sú štandardne zobrazené; v prípade potreby ich zobrazenia je možné program spustiť s parametrom `-f`.

## Impera: náprava chybného delenia

Získané súbory podstatne zredujú množstvo dát pre vizuálnu, prípadne ďalšiu softvérovú kontrolu.

V prípade zistenia výskytu nesprávneho delenia je možné správny spôsob zadať pomocou tabuľky výnimiek `\hyphenation{}`, ktorá sa vzťahuje na aktuálne nastavený jazyk. Pre zle rozdelené slová v tomto článku je možné zadať napríklad `\hyphenation{náj-de zna-mien-ka}`, pokiaľ napriek zásadám v [1] akceptujeme prenášanie dvoch písmen na nový riadok. Ak sa slovo v danom tvare nemá deliť vôbec, treba použiť `\hyphenation{nájde}`.

Minimálne počty písmen, ktoré musia po rozdelení slova zostať na konci riadku, a ktoré musia byť prenesené na nový riadok, je možné nastaviť pomocou číselných registrov `\lefthyphenmin` a `\righthyphenmin`.

Delenie konkrétneho výskytu slova je možné predpísať vsunutím makra `\-` na miesta, v ktorých je riadkový zlom akceptovateľný. V prípade náročnejších požiadaviek je možné použiť všeobecnejšie `\discretionary{}{}{}`.

Alternatívne je možné pre lokálny zákaz delenia slova použiť `\hbox{nájde}`, prípadne `\mbox{nájde}` v  $\text{\LaTeX}$ .

Delenie slov sádzaných konkrétnym fontom `\menofontu` je možné zakázať pomocou nastavenia `\hyphenchar\menofontu=-1`. Úplný zákaz delenia slov je možné dosiahnuť nastavením hodnoty `\hyphenpenalty` a `\exhyphenpenalty` na 10000; pri postupnom znižovaní tejto hodnoty sa zvyšuje frekvencia rozdelených slov. Aktuálna hodnota týchto dvoch parametrov sa vyhodnocuje na konci každého odseku.

Osamotenú predložku je možné pripojiť k slovu známym makrom `~` (vlnovka), a to buď ručne alebo automatizovane napríklad pomocou programu `vlna` [5] alebo (v prípade, že použitý  $\text{\TeX}$  podporuje rozšírenie `enc\TeX`) makier `encxvlna` [7].

Ďalšie možnosti ovplyvnenia delenia slov a riadkového zlomu sú popísané v ← knihách [4, 6].

## Systemové požiadavky

Program je napísaný v jazyku Perl [8, 2], je ho teda možné použiť na každej platforme, pre ktorú je dostupný interpretér Perlu. Pred použitím je potrebné program (súbor `findhyph` v systéme Unix/Linux, `findhyph.bat` v systéme Win- ←  
dows) skopírovať do adresára, ktorý je zahrnutý v systémovej ceste (PATH), prípadne do pracovného adresára. Interpretér Perlu by mal byť nainštalovaný v adresári `/usr/bin/` pod unixovými systémami, alebo kdekoľvek v systémovej ceste pod systémom Windows. Program je voľne dostupný v archíve CTAN.ORG [3].

## Zoznam literatúry

- [1] *Základní pravidla sazby. Oborová norma 88 2503.* [Czech and Slovak Type- ←  
setting Rules.] Praha, Vydavatelství Úřadu pro normalizaci a měření, 1974.
- [2] Bilisoly, Roger. *Practical Text Mining with Perl.* [Praktická extrakcia z textov  
za pomoci jazyka Perl.] Hoboken, New Jersey, USA, John Wiley & Sons, Inc.,  
2008. ISBN 978-0-470-17643-6.
- [3] Budaj, Martin. Program `findhyph`. [The `findhyph` software.] Dostupný z ad- ←  
resára: <http://ctan.org/tex-archive/support/findhyph/>
- [4] Knuth, Donald Ervin. *The T<sub>E</sub>Xbook*. Ch. 14: How T<sub>E</sub>X Breaks Paragraphs into  
Lines. Reading, Massachusetts, Addison-Wesley, 1984. ISBN 0-201-13448-9.
- [5] Olšák, Petr. Program `vlna`. [The `vlna` software.] Dostupný z FTP servera:  
<ftp://math.feld.cvut.cz/pub/olsak/vlna/>
- [6] Olšák, Petr. *T<sub>E</sub>Xbook naruby.* [T<sub>E</sub>Xbook Inside Out.] Kapitola 6.4: Řádkový  
zlom. Brno, Konvoj 1996. ISBN 80-85615-64-9.
- [7] Olšák, Petr; Wagner, Zdeněk. Makrá `encxvlna`. [The `encxvlna` macros.]  
Dostupné z: <http://ctan.org/tex-archive/macros/generic/encxvlna/>
- [8] Wall, Larry; Christiansen, Tom; Schwartz, Randal L. *Programování v jazyce  
Perl.* [Perl Programming.] Praha, Computer Press 1997. ISBN 80-85896-95-8.

## Summary: Divide et impera—The `findhyph` program

The article presents a simple computer program, `findhyph`, which generates a list of all words hyphenated in documents processed by T<sub>E</sub>X. This program can be downloaded from the CTAN.ORG server.

**Keywords:** Plain T<sub>E</sub>X, Perl, Checking hyphenated words, `findhyph` program.

*Martin Budaj, m.b@speleo.sk, C<sub>S</sub>TUG c/o FEL ČVUT  
Technická 2, Praha, CZ-166 27, Czech Republic*