

# Zpravodaj Československého sdružení uživatelů TeXu

---

Martin Pavlík

České třídění / Exkurze do systémů zpracování textu

*Zpravodaj Československého sdružení uživatelů TeXu*, Vol. 6 (1996), No. 1, 31–36

Persistent URL: <http://dml.cz/dmlcz/149751>

## Terms of use:

© Československé sdružení uživatelů TeXu, 1996

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>

*Následující dva texty vznikly jako semestrální práce z předmětu „Publikační systém T<sub>E</sub>X“, který přednáší Petr Olšák na Elektrotechnické fakultě ČVUT. Po drobných, zejména jazykových, úpravách jsou zařazeny do našeho bulletinu, přičemž studenti (autoři jednotlivých příspěvků) s touto formou zveřejnění souhlasí.*

---

---

## České třídění

### Exkurze do systémů zpracování textu

---

MARTIN PAVLÍK

#### Evoluce pokračuje – rodí se počítač

Člověk je tvor společenský a zřejmě z nedostatku osobního kontaktu s ostatními lidmi si vymyslel počítač, aby se necítil tak osamělý. Mnoho programátorů po celém světě tráví dlouhé noci před obrazovkami svých nových přátel a ve snaze vdechnout trochu života elektronickému miláčkoví ho zahrnují kilobajty, v posledních letech spíše megabajty naklofaného textu<sup>1</sup> očekávaje kloudný výsledek. V zápalu nadšení však opomínají býti lidmi a sami se pomalu stávají tvory digitálními bez velké dávky estetického citění. Toto se projevuje zejména v uživatelské nepřátelskosti některých programů, které jsou svými stvořiteli vybaveny mnoha neužitečnými funkcemi, a postrádají to, co je opravdu potřeba.

#### Jazykový babylon

Abychom si vzájemně *lépe* rozuměli, mluvíme na Zemi mnoha různými jazyky. Co národ, to jiná řeč a odlišné zvyklosti. V posledním desetiletí se stále více množí nový digitální tvor a s ním se rodí i nová kultura a nový jazyk. Počítače, bohužel, nejsou tak flexibilní jako lidé a nejsou schopny se přizpůsobit řeči lidské, proto se člověk pozvolna přizpůsobuje počítačům. V našich zeměpisných šířkách vzniká nový jazyk, velmi podobný

---

<sup>1</sup>Slovo *text* se nečte *techt*!

češtině, avšak s mnoha odlišnostmi. Toto *nářečí* se jmenuje „cestina“ a některá jeho pravidla přiblíží následující pojednání. Jak dále uvidíte, „cestina“ není jen jedna, ale liší se počítač od počítače a program od programu.

## Znakové sady

Kolébku počítačů jsou Spojené Státy a protože v této části světa řeč a písmo neprodělaly změny způsobené prací našich jazykotvůrců, jsou i počítače ve své původní podstatě ušetřeny všelijakých dlouhých a krátkých nabodeniček. To je jistě pro polovinu světa velmi příznivé, avšak při expanzi výpočetní  $\text{\TeX}$ niky do naší malé země to s sebou přináší jisté komplikace co se týče znakových sad a ostatních pravopisných odlišností. Na neštěstí problému české znakové sady se ujalo hned několik pohotových tvůrců, každý si šel svou vlastní cestou (zaručeně tou nejlepší) a jako výsledek zde máme několik vzájemně neslučitelných výsledků. V tomto zmatku není mnohdy snadné přesvědčit vlastní tiskárnu aby tiskla česky a ne třeba hebrejsky. Osobně se přiznám, že na počítači píši raději vše anglicky, než abych se potýkal cestinou.

## Třídění dle abecedy

Bohužel, znakovou sadou problémy s implementací češtiny zdaleka nekončí. V textových editorech, tabulkových procesorech a jiných produktech je jako jedna z nepostradatelných funkcí třídění dle abecedy a vytváření různých rejstříků a seznamů. Znaky s akcenty jsou umístěny v horní polovině znakové sady a pořadí znaků v takovéto sadě tedy neodpovídá abecednímu řazení. Třídí-li program jednotlivá hesla podle pořadí znaku ve znakové sadě, dopouští se tak chyb v řazení, které jsou nepřijatelné.

České abecední řazení je specifikováno normou ČSN 01 0181, kde kromě pořadí jednotlivých znaků v abecedě jsou zakotvena také ostatní pravidla, týkající se pořadí velkých a malých písmen, číslic, ostatních znaků, a pravidla pro řazení víceslovných hesel. V této normě je také několik nejasností nebo sporných bodů, které nabízejí více vzájemně odlišných interpretací. Jako nejasné se jeví například třídění téměř shodných slov, lišících se pouze ve znacích s akcenty, nebo shodných slov odlišných pouze v počtu nebo umístění velkých a malých písmen. Norma zde

rozlišuje pořadí na základě počtu velkých nebo akcentovaných písmen, případně podle jejich umístění ve slově. Uvedená pravidla jsou nejasná a často i protichůdná.

Podle mého názoru je vhodné tuto normu přepracovat a použít trochu více zdravého rozumu. Pro snazší orientaci a vyhledávání je vhodné třídit jinak shodná slova postupným porovnáváním znak po znaku zleva doprava, kdy velké písmeno stojí v pořadí za písmenem malým a znak s akcentem za znakem bez akcentu.

## Test vybraných programů

Třídící algoritmy v textových editorech a tabulkových procesorech se liší program od programu a jsou více či méně špatné či dobré. Namátkou jsem vybral několik programových produktů, jejichž třídící algoritmy jsem otestoval a dosažené výsledky předkládám v následujících řádcích.

Byly testovány programy Microsoft Word 5.0 a Claris Works 2.0 na počítači Apple, dále český Word Perfect 5.1 a příkaz Sort operačního systému MS-DOS 6.0. U testu jednotlivých programů jsou v případě (vcelku) správného řazení uváděny pouze odlišnosti.

### *Pravidla správného řazení*

|                |                  |                                                    |
|----------------|------------------|----------------------------------------------------|
| abc nástrojaře | ≺ ABC nástrojaře | (protože abc ≺ ABC)                                |
| abc            | ≺ ABC            | (protože malá písmena ≺ velká)                     |
| ABC            | ≺ abc frézaře    | (protože prázdné slovo je dříve než slovo frézaře) |
| abc frézaře    | ≺ abc nástrojaře | (podle druhého slova)                              |
| abc nástrojaře | ≺ ABC nástrojaře | (protože abc ≺ ABC)                                |
| lalálá         | ≺ lálalá         | (protože la ≺ lá)                                  |
| nadíva         | ≺ nádiva         | (protože na ≺ ná)                                  |
| plagiát        | ≺ plachta        | (protože g ≺ ch)                                   |
| pláně          | ≺ plánička       | (protože é ≺ i)                                    |
| plánička       | ≺ Plánička       | (protože p ≺ P)                                    |
| pláňka         | ≺ plankton       | (protože a ≺ t)                                    |
| plášť          | ≺ plat           | (protože š ≺ t)                                    |

|        |            |                             |
|--------|------------|-----------------------------|
| plat   | ≺ plát     | (protože a ≺ á)             |
| platno | ≺ plátno   | (protože a ≺ á)             |
| plátno | ≺ platnost | (protože platno ≺ platnost) |
| sténá  | ≺ stěna    | (protože é ≺ ě)             |

---

*Vzor správného řazení*

abc  
ABC  
abc frézaře  
abc nástrojaře  
ABC nástrojaře  
lalálá  
lálalá  
nadívá  
nádiva  
plagiát  
plachta  
pláně  
plánička  
Plánička  
plaňka  
plankton  
plášť  
plat  
plát  
platno  
plátno  
platnost  
sténá  
stěna

abc nástrojaře

ABC

ABC nástrojaře

*Řazení programem*

*Claris Works 2.0*

abc

ABC

abc frézaře

abc nástrojaře

ABC nástrojaře

*Řazení programem*

*Word Perfect 5.1 (česká verze)*

abc nástrojaře

abc frézaře

abc

ABC nástrojaře

ABC

*Řazení programem*

*Word Perfect 6.0 (anglická verze)*

abc nástrojaře

abc

abc frézaře

ABC

*Řazení programem*

*Microsoft Word 5.0*

abc

abc frézaře

ABC nástrojaře  
plaňka  
plachta  
plášť  
plagiát  
plánička  
Plánička  
plankton  
pláně  
plat

*Řazení programem*  
*Sort DOS 6.0 (COUNTRY=042)*

ABC  
abc

abc frézaře  
abc nástrojaře  
ABC nástrojaře  
plaňka  
plachta  
plagiát  
plankton  
plat  
platno  
platnost  
plášť  
plánička  
Plánička  
pláně  
plát  
plátno

## Závěry testu

Program CSR [1] řadí podle uvedeného vzoru. Textový editor Microsoft Word 5.0 uspěl vcelku dobře, jen při rozlišování pořadí velkých a malých písmen ve víceslovných heslech měl problémy. Shodná hesla jsou tříděna nejprve podle prvního slova, kde malá písmena mají správně přednost před velkými. Tabulkový procesor Claris Works seřadil všechna hesla bez jediné chyby, i víceslovná hesla jsou seřazena správně nejprve podle abecedy a až poté podle velkých a malých písmen. Textový editor Word Perfect 5.1 pod DOSem řadil hesla správně, jen víceslovná hesla jsou seřazena poněkud podivně. Zdá se, jakoby algoritmus řadil podle druhého slova v heslu spíše sestupně než vzestupně. Řazení probíhá nejprve podle velkých písmen v prvním hesle, až poté podle druhého hesla. Textový editor Word Perfect 6.0 v anglické verzi není upraven pro třídění českých textů, výsledek tomu odpovídá. Pouze se zdá nepochopitelné, jak se slovo plaňka dostalo před slovo plachta. Je také patrné, že třídění probíhá pouze podle prvního slova v heslu. Příkaz Sort operačního systému DOS je velmi neschopný. Nedokáže rozeznat velká a malá písmena a jeho práce s akcenty je též na nic. Možná by se choval lépe, kdyby bylo použito kódování znaků podle MicroSoftu namísto

Kamenických. Toto kódování však pro jeho nedostatky téměř nikdo nepoužívá<sup>2</sup>.

## Názor autora

Drobné nedostatky v řazení, jako třeba prioritní řazení podle velkých písmen prvního slova před druhým slovem hesla, se dají v některých případech opomenout nebo obejít (lze si na ně zvyknout). Nesprávné řazení akcentů je však nepříjemné a proto nepřipustné.

Je třeba si však uvědomit, že vlastní algoritmus řazení není vše, pokud nejsou uživateli poskytnuty ostatní potřebné funkce jako vytváření indexů, rejstříků, a jejich vzájemná provázanost. Program CSR řadí správně, ale je neohrabaný a chybí jeho provázanost s ostatními funkcemi. Velmi nepříjemné je omezení řazení na takzvané volné řádky vstupního souboru.<sup>3</sup> Bylo by přinejmenším vhodné implementovat klíčová slova pro zahájení a ukončení bloku textu, ve kterém je třídění požadováno, např. `\csrbegin` a `\csrend`. Parametr třídění, tj. rozsah sloupců podle kterých se třídí, by měl také být redefinovatelný např. příkazem `\csrpara.....`

## Literatura

[1] Petr Olšák, *Program csr (Czech Sort) – abecední řazení podle normy*. Zpravodaj ČSTUGu, 3:126–139, 1994.

---

<sup>2</sup>Kódová stránka 852 podle MicroSoftu a IBM je jediným standardem pro češtinu v OS/2, žádný jiný kód se zde nepoužívá. Pozn. red.

<sup>3</sup>Program CSR rozlišuje pevné a volné řádky vstupního souboru. Pevné řádky jsou všechny komentářové řádky od začátku souboru, dále následuje pole volných řádků a první následující komentářový řádek pole volných řádků opět uzavře.