Zpravodaj Československého sdružení uživatelů TeXu

Tomáš Hála; Gita Urbanová

Softwarová podpora korektur interpunkce

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 22 (2012), No. 2, 120–126

Persistent URL: http://dml.cz/dmlcz/149962

Softwarová podpora korektur interpunkce

Tomáš Hála, Gita Urbanová

Abstrakt:

Článek se zabývá možnostmi automatizace korektur interpunkčních jevů. Nejprve byly srovnány možnosti existujícího programového vybavení. Dále byla stanovena míra chybovosti při psaní interpunkčních jevů (2,35 % pro ekonomické texty, 0,96 % pro texty z oboru informatika). Celkově vysoká hodnota vedla k rozhodnutí implementovat jednoduchý korektor těchto jevů. Tomu předcházelo určení množiny jevů vhodných pro automatickou korekci a množiny jevů, které lze detekovat, ale již ne jednoduše automatizovaně opravit.

Klíčová slova: korektura, interpunkční znaménka, písařské chyby, zjištění chybovosti, automatická oprava, detekce nejednoznačných jevů

Úvod

Korektura patří mezi neodmyslitelné součásti procesu vzniku kvalitní tiskoviny či elektronické prezentace. Jejím smyslem je odhalení pravopisných, typografických, případně i věcných chyb v dokumentu.

Jednou z oblastí, ve které se chybuje, je interpunkce. Chyby v psaní interpunkčních znamének mohou být způsobeny jak neznalostí autora, tak – což je častější – nepozorností při sestavování textu.

Před předáním na korektury je vždy lepší se pokusit odstranit co nejvíce případných chyb automatizovaným způsobem, aby se korektor-člověk mohl soustředit na ostatní jevy.

Proto jsme se rozhodli prozkoumat jednak možnosti automatizace korektur interpunkčních jevů v existujícím programovém vybavení, jednak určit míru chybovosti, která by posloužila k rozhodnutí, zda se vyplatí korektor implementovat.

Problémy s psaním interpunkce nebo s překlepy v interpunkci se objevují nejen v obyčejných dokumentech, ale také v pracích vědeckých, o což jsme se opřeli v další úvaze.

Programové vybavení pro korektury textu

Automatizované korekce provádí programy nazývané textové korektory. Ty se vyskytují jako součást textového editoru či procesoru, či jako samostatné programy pro korekci dokumentu.

 doi: 10.5300/2012-2/120

Textové procesory v kancelářských balících obsahují rozsáhlé možnosti a zjednodušení při pořizování textového dokumentu. Například korektor implementovaný v programu MS Word se člení do několika samostatných oddílů – automatické opravy textu, kontrola pravopisu, kontrola gramatiky a nástroj Tezaurus, který napomáhá nalezení správného slova. (Mansfield, 1998) Opravami interpunkce se přímo nezabývá, má však implicitně aktivovány některé další funkce, mezi něž patří například změna na velká písmena na začátku vět, což se ovšem může negativně promítnout v situacích, kdy tečka neoznačuje konec věty ale zkratku či iniciály. Tento repertoár obsahují všechny verze od „97“ do současnosti. Dále je zde implementována opravu počátečník uvozovek ("→„).

Kancelářský balík LibreOffice (2010) nabízí v rámci automatických oprav odstranění dvojitých mezer, náhradu spojovníku a sekvence dvou spojovníků za půlčtverčíkovou pomlčku (též v MS Word) a rovněž odstranění mezer a tabulátorů na začátcích a koncích řádků a odstavců. Bez nastavování nahrazuje sekvenci tří teček výpustkem. Dalšími jevy z oblasti interpunkce či sazby značek se nezabývá.

V kancelářských balících lze dále využít funkci automatických náhrad, ty si však uživatel musí definovat sám.

Unixové programy, sloužící často k psaní a úpravě textu (`joe`, `vim` apod.), byly v souladu se základní unixovou zásadou vývoje OS Unix (*psát programy, které budou dělat právě jednu věc, a tu budou dělat dobře*; Brandejs, 2003) konstruovány pouze pro tyto editační účely a další funkce neobsahují.

Příkladem programu ctícího tuto zásadu může být program `vlna` (Olšák, 2002, 2010). Jedná se o jednoúčelový malý program, který v dokumentu vyhledá všechny jednopísmenné předložky a spojky, za něž vloží místo obyčejné mezery mezeru nezlomitelnou.

Ze stejného důvodu jsou v OS Unix opravy řešeny samostatným programem `aspell`, korektor však je určen pro opravy překlepů, interpunkcí se stejně jako dosud zmíněné programy nezabývá.

Jediným programem, který sloužil v minulosti ke sledovanému účelu, byl program `ikor` (Rybička, [1997]). Jednalo se o interaktivní korektor písařských chyb. Korektor byl založen na lexikální analýze vybraných atomů a na stanovení jejich pořadí, které bylo podrobeno analýze vhodnosti. Písařskou chybou se zde rozumělo nesprávné pořadí interpunkce a mezerování. Zjištěné nesrovnalosti program zapisoval do změnového souboru, z něhož se ve druhé fázi čerpaly informace pro vlastní korekci textu. Program byl vytvořen pro OS DOS a není veřejně dostupný.

Za zmínku stojí také program `lacheck`, určený k předběžné kontrole dokumentů napsaných v systému LaTeX. I přesto, že jeho nejvýznamější praktický přínos spočívá v kontrole párovosti závorek všech druhů včetně kontroly párů příkazů pro sázecí prostředí, je schopen odhalit nesprávné mezerování, doporučit použití výpustků apod. Jedná se však pouze o detekci velmi malé množiny jevů. Kromě toho jeho užití pro jiné formáty je dosti omezené.

Interpunkcí se zabývá také program Grammaticon (Lingea, 2012). Jedná se o komplexnější, interaktivní řešení odhalující – kromě řady gramatických jevů – chybnou interpunkci a chybějící nebo nadbytečné čárky ve větách. Program Grammaticon však bohužel patří mezi komerční produkty.

Shrneme-li situaci ve zkoumaném programovém vybavení, musíme konstatovat, že neexistuje žádný produkt, který by se zabýval komplexně detekcí a případnou korekcí písařských chyb a současně nebyl (výlučně) interaktivní, dále aby byl dostupný a volně šiřitelný.

Hodnocení chybovosti v odborných textech

Na základě empirických zkušeností se sazbou a korekturami textu bylo zřejmé, že určité množství chyb je způsobnosti písařskými nepřesnostmi v oblasti interpunkčních znamének. Bylo však potřeba toto množství kvantifikovat, aby bylo možné rozhodnout, zda se vyplatí korektor implementovat.

Pro posouzení chybování byl použit náhodný výběr původních vědeckých prací ekonomických oborů (21 textů) a vědeckých pracích oboru informatika (16 textů). Oba obory byly hodnoceny odděleně. Ve všech případech se jednalo o rukopis, tedy ještě nekorigovaný dokument.

Každá práce byla hodnocena týmž způsobem, zpracován je pouze český nebo slovenský text. Vložené grafy, tabulky, obrázky a jejich popisky stejně jako závěrečné autorovy osobní údaje nebyly pro statistiku chybování v interpunkci uvažovány. Slovenské texty jsme z hodnocení záměrně nevyloučili, neboť pravidla slovenské interpunkce jsou téměř shodná s českými.

V každém textu byly zvlášť posouzeny jednotlivé jevy. Podrobné tabulky četností členěné podle souborů a jevů prezentovala Urbanová (2003). Zde uvádíme pouze souhrnné výsledky.

Součástí kvalitního zpracování dokumentu by měly být nástroje, které spolehlivě eliminují výskyt chyby, jak gramatické, tak případně typografické či jiné. Každá odborná práce by ve své výsledné podobě obsahovat chyby neměla.

Na výsledné hodnotě statistiky původních vědeckých prací ekonomických je však znát, že takové bezchybné úrovně zdaleka nedosahují. Kontrola chybovosti v těchto textech byla provedena pouze pro interpunkční znaménka. Podíl chybně zapsaných interpunkčních jevů dosáhl u ekonomických textů 2,35 %, což je na tak odborný materiál dosti značné, a lze soudit, že nástroj k odstranění nebo alespoň snížení procenta chyb by v takovém případě měl své opodstatnění.

Statistické zpracování souboru 16 náhodně zvolených prací v oboru informatika ukázalo, že procentní chybovost v interpunkci činila celkově 0,96 %, což je lepší hodnota než u souboru prací ekonomických. Při kontrole chybování v interpunkci byla aplikována stejná pravidla jako pro práce ekonomické.

Za chyby byla považována znaménka chybně oddělená od ostatního textu, také znaménka chybějící ve smyslu uvozovek či závorek, které se musí vysky-

	obory ekonomické			*obor informatika*		
jev	výskytů celkem	četnost chyb	chybovost [%]	výskytů celkem	četnost chyb	chybovost [%]
tečka	3 143	129	4,10	1 728	22	1,27
čárka	3 125	21	0,67	1 862	12	0,64
uvozovky	212	4	1,89	154	2	1,30
pomlčka	392	5	1,28	284	1	0,35
závorka	816	20	2,45	588	3	0,51
výpustek	12	2	16,67	11	2	18,18
dvojtečka	400	10	2,50	224	4	1,79
středník	29	0	—	18	0	—
otazník	12	0	—	19	1	5,26
vykřičník	0	—	—	0	—	—
odsuvník	0	—	—	0	—	—
celkem	8 141	191	2,35	4 888	47	0,96

Tabulka 1: Hodnocení chybování v interpunkci

tovat v páru. Znaménka, která chybí, například z důvodu neznalosti českého pravopisu nebo pouhého překlepu uvažována nejsou, neboť zkoumán byl pouze zápis interpunkce, nikoliv její správné jazykové použití.

Ve sledovaných pracích se vyskytovala i chybně psaná sousedící interpunkční znaménka. V takovém případě bylo nutné rozhodnout, kterému jevu tato chyba náleží. Například ve spojení: „zima.(chladno)“ chybí mezi tečkou a závorkou mezera. Je tato chyba způsobena chybějící mezerou za tečkou, nebo mezera chybí před závorkou? Chybné psaní bylo správně připsáno závorce, neboť tečka ve své definici obsahuje, že mezerou je zprava oddělována, jestliže se za ní nevyskytuje jiné interpunkční znaménko. Naproti tomu před otevírající závorkou se mezera vyskytuje vždy, pokud neleží na počátku řádku.

Sestavení množiny jevů vhodných pro automatickou korekci a pro detekci

Uvedená interpunkční znaménka spolu s jejich pravidly pro psaní dělíme na dvě skupiny: Znaménka, která lze řešit automatickou korekcí a znaménka, na jejichž chybný zápis reagují varovná hlášení.

Automatická korekce je vhodná pro znaménka, která neobsahují nejednoznačná a na kontextu závislá pravidla pro psaní interpunkce. Patří mezi ně tečka, před kterou se v českém jazyce nikdy nevyskytuje mezera a za níž se mezera naopak povinně vkládá, jestliže nenásleduje jiná interpunkce. Takto automatizovaně lze korektury provést také u středníku, vykřičníku, otazníku, závorek, odsuvníku a uvozovek.

Dvojtečka je automaticky řešena jen ve vztahu k písmenům abecedy, neboť u dvojtečky mezi čísly dochází k jisté nejednoznačnosti. Nevíme totiž, zda se jedná o matematické dělení, nebo číselný poměr (odhlížíme od poměru vyjádřeného písmeny, např. a:b). Stejně tak u čárky není jasné, zda se jedná o výčet čísel oddělený čárkami, nebo o desetinné číslo. Také zde korektor rozhoduje automatizovaně jen v okolí písmen české abecedy.

Jako další případ lze uvést apostrof, vyznačující v lingvistických textech palatizovanou souhlásku. Proto může tudíž stát i na začátcích slov, např. 'sloboda (Pravidlá slovenského pravopisu, 1991). To platí i pro češtinu, přestože Pravidla českého pravopisu tento případ výslovně nezmiňují.

Taktéž lomítko píšeme podle kontextu s mezerami (verše), bez mezer (např. vyjádření alternativ: ano/ne) nebo „střídavě“ /náhrada závorek/.

Pouze *detekce* je aplikována na interpunkci, která se řídí pravidly závislými na kontextu. K takovým jevům náleží výpustek, lomítko a pomlčka. Součástí detekce jsou varovná hlášení, upozorňující například na lichý počet uvozovek v souboru nebo na zápis více desetinných čárek mezi čísly.

Implementace korektoru české interpunkce

Korektor české interpunkce (Urbanová, 2003) byl sestaven v jazyce Perl s využitím regulárních výrazů, které jsou účinným nástrojem pro zpracování řetězců v textových souborech. Proto se pomocí nich velmi dobře vyhledávají interpunkční znaménka a zkoumají korektní zápisy identifikovaných jevů.

Každý jev obsluhovaný tímto korektorem interpunkce je implementován jako samostatná procedura, aby bylo možné volit potřebné jevy a případně jevy nežádoucí vypustit. Příkladem nežádoucí korekce může být například úprava u kulatých závorek, užívané jako součást chemických vzorců, u komplexních sloučenin je pak nežádoucí i korekce hranatých závorek.

Jako další případ lze uvést apostrof, vyznačující v lingvistických textech palatizovanou souhlásku. Proto může tudíž stát i na začátcích slov, např. 'sloboda (Pravidlá slovenského pravopisu, 1991). To platí i pro češtinu, přestože Pravidla českého pravopisu tento případ výslovně nezmiňují.

Taktéž lomítko píšeme podle kontextu s mezerami (verše), bez mezer (např. vyjádření alternativ: ano/ne) nebo „střídavě“ /náhrada závorek/.

Řízení korekce

Korektor realizuje opravy výše uvedených jevů a umožňuje řídit působení jejich korekce pomocí direktiv v textu. Jevy zahrnuté v korektoru jsou řízeny procedurami samostatně, aby bylo možné přidělit každému jevu vlastní direktivu.

Tento přístup byl inspirován programem `vlna` (Olšák, 1995–2010), kde vkládání nezlomitelných mezer lze vypnout direktivou `%~-` a zapnout analogicky

direktiva	význam	direktiva	význam
T	tečka	H	závorky hranaté
C	čárka	K	závorky kulaté
S	středník	Z	závorky složené
D	dvojtečka		
O	otazník	M	redukce mezer
V	vykřičník	L	lomítko
U	uvozovky		
A	apostrof	E	zahrnuje všechny opravy

Tabulka 2: Seznam direktiv pro řízení korektury interpunkce

pomocí `%~+`. V nežádoucích místech, například v matematických prostředích či v prostředí `verbatim`, se tak automatické opravy neprovádějí a text zůstane nedotčen.

Pro tento korektor jsme jako prefix zvolili `%!`, za nímž následuje jednopísmenná direktiva se znaménkem `+` nebo `-`. Seznam direktiv, obsažený též v nápovědě k programu, uvádí tabulka 2.

Požadované nastavení je také možné načíst z konfiguračního souboru.

Varovná hlášení
Snahou korektoru je také v některých případech, kde nelze u nalezených jevů jednoznačně rozhodnout, alespoň vypsat varovná hlášení. Hlášení vypisuje korektor v těchto případech:

- v souboru objevil lichý počet uvozovek;
- konfigurační soubor obsahuje neznámý symbol (pro jistotu následuje volání funkce `die`);
- v souboru se vyskytla pomlčka ohraničená mezerou jen z jedné strany;
- v souboru se vyskytují dvě nebo čtyři po sobě jdoucí tečky;
- soubor obsahuje chybně zapsané desetinné číslo, např.: 1,223,4.

Závěr

Automatizované korektury přináší uživateli mnohá zjednodušení a krácení času stráveného zpracováním textů. Většina textových editorů či procesorů obsahuje korektor pravopisu, některé i korektor gramatiky, nikde však se nenachází k dispozici nástroj k opravě písařských chyb v interpunkci. Ani nástroj automatických oprav uživateli neposlouží, neboť by uživatel si musel všechny sledované jevy či chyby sám vložit.

Sledováním četnosti různých typů chyb způsobených špatným psaním interpunkce bylo zjištěno, že písařské chyby v oblasti interpunkce nejsou zdaleka zanedbatelné. Proto byl implementován jednoduchý korektor.

Představená verze korektoru české interpunkce uživateli zcela poslouží v případě zpracování holého textu. Zpracování dokumentů značkovaných v imple-

mentacích jazyka TEX vyžaduje potlačení oprav okolo složených závorek. Do budoucna je možné uvažovat rozšíření na další formáty, např. HTML.

Při zpracování odborných dokumentů je však třeba mít na paměti, že automatizované opravy některých znamének by mohly být spíše na škodu, některé případy již byly zmíněny.

Korektor byl vytvořen primárně pro optimalizaci zpracování textů v češtině, lze jej však využít i pro zpracování textů ve slovenštině, neboť pravidla psaní interpunkčních znamének se v těchto dvou jazycích neliší.

Program se nachází na adrese `www.korektury.cz/software/punch` a je volně ke stažení.

Reference

Brandejs, Michal. *Linux : Praktický průvodce.* 2. vyd. Brno : Konvoj, 2003. 312 s. ISBN 80-7302-050-5.

Grammaticon [on-line]. [s. a.]. [cit. 2012-06-25]. Dostupné na: `http://www.lingea.cz/grammaticon.htm`.

lacheck v. 1.26 [software]. 1998.

LibreOffice 3.3 [software]. c2000, 2010.

Mansfield, Ron. *Word 97.* 1. vyd. Grada Publishing : Praha, 1998. 887 s. ISBN 80-7169-517-3.

Olšák, Petr. *vlna* v. 1.2, 1.5 [software]. 2002, 2010.

Pravidlá slovenského pravopisu. 1. vyd. Bratislava : Slovenská akadémia vied, 1991. 536 s. ISBN 80-224-0080-7.

Rybička, Jiří. *ikor* [software]. [1997].

Urbanová, Gita. *Implementace korektoru české interpunkce.* (Bakalářská práce.) Brno : Mendelova zemědělská a lesnická univerzita, 2003. 35 s.

Summary: Software Support for Punctuation Proof

This paper is focused on possibilities of automating corrections of punctuation mistypings. First, functionality of current software tools is compared. Second, the error rate of punctuation mistypings is determined (2.35% and 0.96% for economics and informatics texts, respectively). These high values are a compelling reason to implement a simple program for correcting such mistypings. Also, we identify mistypings suitable for automatic corrections as well as those that can be detected but are not easy to automatically correct.

Key words: proof, punctuation marks, mistyping, error rate, automated correction, detection of ambiguous phenomena

thala@mendelu.cz

Mendelova univerzita v Brně, Provozně ekonomická fakulta, ústav informatiky, Zemědělská 1, CZ 613 00 Brno