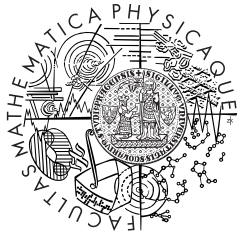
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Drahomíra Doležalová

Počítačový modul stylistického korektoru češtiny

Ústav formální a aplikované lingvistiky
Vedoucí diplomové práce: prof. PhDr. Jarmila Panevová, DrSc.
Studijní program: Informatika

Děkuji prof. Jarmile Panevové za obětavé vedení této práce, v duchu nedávného výroku k našemu nejmenovanému kolegovi „že vy už jí jednou nedáte do zubů“ se divím, že mi do nich někdy nedala sama. Dále děkuji Janu Hajičovi a Pavlovi Květoňovi, bez jejichž softwarových nástrojů (morfologie a programovací jazyk LanGR) by byla praktická část této práce stěží realizovatelná, Markétě Lopatkové, Karlovi Olivovi, Mileně Hnátkové, Tomášovi Holanovi a Jarce Hlaváčové, jakož i Mirkovi Spoustovi a Martinovi Pergelovi za všeliké rady a připomínky k teoretické části a Josefovi Poláchovi za pomoc se sazbou. Děkuji též ÚČNK za poskytnutí přístupu ke korpusu a firmě Internet Info za poskytnutí dat.

Největší dík patří (proti jeho vůli) Nikimu Petkevičovi nejen za spoustu nápadů, rad, připomínek a postrčení správným směrem, ale i za jeho nezdolné nadšení a optimismus, kterým umí tak dobře nakazit i ostatní.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím uvedených pramenů. Souhlasím se zapůjčováním práce.

V Praze 12. dubna 2004

Drahomíra Doležalová

Obsah

Název práce: Počítačový modul stylistického korektoru češtiny
Autor: Drahomíra Doležalová
Katedra (ústav): Ústav formální a aplikované lingvistiky
Vedoucí diplomové práce: prof. PhDr. Jarmila Panevová, DrSc.
e-mail vedoucí: panevova@ufal.ms.mff.cuni.cz
Abstrakt: Věnujeme se vybraným problémům na pomezí stylistiky a gramatiky (například atrakce, stupňování, příslovečné spřežky, rozlišení významu spojky *nebo*, kontaminace, vokalizace předložek). Teoretická část práce obsahuje podrobný rozbor všech zkoumaných jevů, z větší části původní, zejména s ohledem na možnost formálního popisu těchto jevů a na jejich automatickou odhalitelnost. Za nejcennější výsledek této části práce považujeme rozbor atrakcí. Praktická část obsahuje implementaci v jazyce LanGR. Zformulovali a implementovali jsme pravidla, která vyhledávají některé ze zkoumaných jevů v morfologicky označkovaném (obecně nedisambiguovaném) textu. Provedli jsme testování na PDT (1,5 milionu ručně disambiguovaných slov), kde jsme při zachování 100% přesnosti odhalili 59 stylistických chyb. Implementace dále zahrnuje skript pro vokalizaci předložek napsaný v programovacím jazyce Perl.
Klíčová slova: stylistika, čeština, pravidlové metody

Title: A Style Checker for Czech
Author: Drahomíra Doležalová
Department: Institute of Formal and Applied Linguistics
Supervisor: prof. PhDr. Jarmila Panevová, DrSc.
Supervisor's e-mail address: panevova@ufal.ms.mff.cuni.cz
Abstract: We have studied selected stylistic and grammatic problems of Czech (e. g. attraction, degrees of comparison, compound adverbs, conjunction *nebo* meaning determination, crossing of government, vocalization of prepositions). Theoretical part of our work consists of (mostly original) detailed analyses of all the phenomena studied; these analyses focus on the possibilities of the formal description of the phenonema and their automatic detection. We consider the results connected with an analysis of attraction the most valuable asset of this part. The implementation part of the work contains several rules written in the Language for Grammatical Rules - LanGR. They are able to find some problematic constructions in morphologically annotated (in principle nondisambiguated) text. We performed tests on the Prague Dependency Tree corpus (PDT) (1 500 000 manually disambiguated words), where we succeeded in finding 59 style mistakes keeping absolute precision. The implementation contains also a Perl script for vocalization of prepositions in Czech.
Keywords: style, Czech language, rule-based methods

Úvod

Tématem této práce jsou některé jazykové jevy, které bývají zdrojem víceznačností a stylistických neobratností.

V první části uvádíme několik motivačních příkladů a zasazujeme výzkum do praktického rámce – popisujeme, odkud příklady pocházejí a čeho bylo na tomto poli dosud dosaženo.

Ve druhé části podrobně teoreticky rozebíráme jednotlivé zkoumané jevy, jejich varianty, rozdělujeme je podle hloubky úrovně jazykového popisu, do které zasahují, a podle toho, zda a proč je lze nebo nelze automaticky odhalit. Přiblížíme si také některé obecnější pojmy, které při popisu jevů používáme.

Ve třetí části popisujeme dostupné nástroje morfologické a syntaktické analýzy a automatického zpracování češtiny vůbec, zejména programovací jazyk *LanGR* Pavla Květoně [11], pomocí něhož byla implementace některých pravidel provedena, spolu s projektem pravidlové disambiguace a souvisejícího gramatického korektoru [13], který je s naší prací úzce svázán.

Ve čtvrté části se omezujeme na ty z dříve představených jazykových jevů, které lze alespoň částečně automaticky odhalit či dokonce opravit. Popíšeme pravidla, která byla za tímto účelem implementována v jazyce LanGR či jiným způsobem.

V závěrečné části hodnotíme dosaženou úspěšnost a rozebíráme možnosti vylepšení.

Kapitola 1

Motivace

Jarmila vždycky mi radila, abych pravidla syntaxe dodržel. . .

parafráze tvorby Pavla Dobeše

Když se řekne stylistický korektor, většina lidí si představí něco, co po nich bude opravovat velká písmena u oslovení a případně je upozorní na to, že používají v textu příliš dlouhé či krátké věty. V duchu těchto předsudků se bohužel chovají i softwarové produkty, které byly pro češtinu dosud vyvinuty, konkrétně jediný, „gramaticko-stylistický" korektor firmy Lingea, který lze zakoupit a nainstalovat do textového editoru Microsoft Word. Tento korektor upozorňuje například na příliš krátké věty, rozkazovací způsob (a to vždy, i v případě, že byl mylně pochopen vinou chybné morfologické disambiguace), velká a malá písmena u oslovování či příliš mnoho částic („příliš mnoho" je pro něj i jedna jediná).

Je tedy zřejmé, že tento produkt není zcela vyhovující, ať už z důvodu (místy zcela zbytečných) implementačních chyb, nebo proto, že lze samořejmě zajít mnohem dále – slovosledem počínaje a víceznačnostmi hraničícími s gramatickou chybou konče.

Při zadávání této práce již byly doporučeny některé konkrétní jevy ke zkoumání spolu s literaturou, která se k nim váže. Nakonec jsme však většinu těchto příkladů zavrhli a zabýváme se téměř výhradně jevy vypozorovanými z vlastní zkušenosti a praxe a také pravidelně opravovanými při korekturách odborných počítačových článků pro server Root.cz [3] a další média. Z těchto článků také pochází většina příkladů, jsou to všechny, u kterých není uveden zdroj (*Měšec* a *Lupa* jsou další zpravodajské servery téhož provozovatele).

Co se týče doporučených jazykových jevů, od prof. Jarmily Panevové pocházejí levá adjektivní rozvití substantiva a od *atraktivního* doc. Vladimíra Petkeviče atrakce, stupňování adjektiv a adverbií a vokalizace předložek, v neposlední řadě také mnoho cenných rad ke všem dalším kapitolám.

Osobně sem neznal Komenskýho, ale řek bych, že Niki ho převálcuje.

„nejmenovaný správce sítě" ústy Karla Kučery [10]

Kapitola 2

Teoretický rozbor

V této části práce rozebíráme jednotlivé jazykové jevy, které jsou předmětem naší analýzy, z hlediska teoretického – jak se dělí, proč vznikají, kde končí neobratnost a začíná chyba apod. Orientačně si řekneme, zda máme naději jev odhalit, a v případě, že ano, budou podrobnosti následovat v příslušné implementační kapitole. Je-li jev automaticky neodhalitelný, ukážeme si proč.

Při automatickém odhalování chyb a výpisu varování se snažíme držet přesnost (precision) blízkou 100 %, nejen z toho důvodu, že pro získání důvěry uživatele je záhodno nevarovat jej zbytečně, ale také proto, že je to v souladu s pojetím celého projektu pravidlové disambiguace a souvisejícího gramatického korektoru. Proto jsou některá pravidla možná až příliš opatrná, čímž přicházíme o mnoho (místy i velmi pravděpodobných) případů chyb, zato si však můžeme být téměř jisti, že všechny chyby námi oznámené jsou chybami skutečnými.

Předchozí odstavec se sice již týká převážně praktických otázek implementace, je však třeba mít jej neustále na paměti už při čtení teoretických rozborů. Teprve pak je totiž jasné, proč jsme se rozhodovali tak či onak a proč jsme implementaci řešení některých typů problémů rovnou zavrhli.

Odkazujeme-li se na výskyty v ČNK [1], míníme tím Korpus synchronní češtiny SYN2000.

2.1 Atrakce

Česká mluvnice [8] uvádí tuto definici: „Atrakce (větná spodoba) záleží v tom, že slovo ve větě je v jiném tvaru, než který mu podle mluvnické závislosti náleží, a to působením slova jiného."

Některé případy jsou již v jazyce ustáleny, a jsou proto správné (*širší než delší* (místo . . . než dlouhý), *vezmi kde vezmi* (místo . . . kde vezmeš)), jiné jsou dosud pokládány za chybné (*ve většině případech*, *ke spoustě lidem*), a právě jejich odhalováním se zde zabýváme.

Následuje rozbor případů podle toho, jaký slovní druh je zasažen a kterých pádů se situace týká.

2.1.1 Substantivum ovlivněné levým rozvitím

Zvrhnutí genitivu v lokál

K této chybě dochází u (převážně deverbativních) substantiv vzoru *stavení* v plurálu, a to vlivem rozvití genitivním adjektivem, které má i lokálové čtení, což pisatele zmate. Naštěstí zde máme značnou naději na odhalení, neboť lokál musí mít předložku. Pokud ji nenalezneme (přímo před řetězcem slov buď s možným lokálovým čtením, či slov nesklonných, např. spojek, viz příklad 1), je tvar substantiva špatně. Také výskyt předložky, která se s lokálem nepojí, nás ujistí o chybě.

Příklad 2.1–1: *Ve skutečnosti jde o povrchní popis života reálných postav plný historických chyb a* ***nepodložených tvrzeních*** *o jejich životech. (recenze filmu Frida, dokina.cz)*

Příklad 2.1–2: *. . . bude řešena možností přidání* ***řady užitečných rozšířeních****.*

Příklad 2.1–3: *. . . také* ***u hardwarových řešeních****.*

Příklad 2.1–4: *Nejprve zjistíme aktuální hodnotu příznaku* ***u zadaných síťových rozhraních****.*

Zvrhnutí lokálu v instrumentál

Jev opět postihuje převážně deverbativní substantiva vzoru *stavení* (tentokrát v singuláru), u nichž vlivem toho, že rozvíjející adjektivum vzoru *jarní* má stejný tvar v lokálu i v instrumentálu, dochází k chybnému výběru pádu (který je pak v rozporu s lokálovou předložkou předcházející celé rozvití). Opět nám zde napomůže fakt, že předložka je obligatorní, tentokrát ke zvýšení úplnosti (recall). Předložku totiž najdeme téměř vždy, čímž budeme upozorněni na chybu.

Nevyskytuje-li se taková předložka před podezřelou sekvencí, předpokládáme, že tato sekvence je správná.

Příklad 2.1–5: *. . . zejména **při větším zatížením** systému*

2.1.2 Adjektivum ovlivněné rozvíjeným substantivem

Tyto případy jsou způsobeny homonymií genitivu a lokálu u mužských substantiv vzoru *hrad* v singuláru, která může ovlivnit pisatele při formulování následujícího pravého adjektivního rozvití.

Zvrhnutí genitivu v lokál

Díky obligatornosti předložky u lokálu je zde naštěstí odhalitelnost opět snadná – nenalezneme-li (před příslušným homonymním tvarem substantiva a případným rozvolněním, viz výše) předložku pojící se s lokálem, je rozvití chybné, případná genitivní předložka nás o tom ujistí.

Příklad 2.1–6: *. . . mají za úkol jediné: rozšiřování a popularizaci Mozilly, výborného **webového browseru založeném** na renderovacím enginu Gecko.*

Příklad 2.1–7: *. . . jsme se opět setkali **u našeho občasníku věnovaném**. . .*

Příklad 2.1–8: *. . . **z modulu obsaženém** v jiném balíčku.*

Zvrhnutí lokálu v genitiv

V tomto případě bohužel obecné odhalení není v našich silách – lokálová předložka nám již nepomůže, neboť je „uspokojena“ rozvíjeným substantivem, a přestože nás takto o lokálovém čtení onoho substantiva ujišťuje, většinou nemůžeme rozhodnout, zda následující adjektivum s genitivním čtením rozvíjí toto substantivum, nebo jiný větný člen (vpředu či vzadu). Problém valence se týká sémantiky a se současnými možnostmi je neřešitelný.

Příklad 2.1–10: *PVM vzniklo z hladu **po systému zastřešujícího** heterogenní síťové prostředí.*

Příklad 2.1–11: *Funkce DiffWDays vrací počet pracovních dnů **v intervalu určeného** parametry.*

Příklad 2.1–12: *Jednoduché použití bodového světla je ukázáno v prvním **demonstračním příkladu uvedeného** na konci tohoto textu.*

2.1.3 Špatné skloňování po kvantifikátoru

Při užití substantiva označujícího množství (*řada*, *většina*, *stovky*...) má následující substantivum tendenci v některých případech přejímat pád substantiva kvantifikačního (typicky lokál, příklady 13 a 14, případně dativ, příklady 15 a 16, či instrumentál, příklad 17), místo aby vyhovělo jeho genitivní valenci.

Příklad 2.1–13: ***Ve většině případech*** *se mi dostalo odpovědi typu:...*

Příklad 2.1–14: *Funkce implementované v těchto knihovnách jsou dnes podporovány* ***na většině platformách****.*

Příklad 2.1–15: *...přístup* ***k většině funkcím*** *ze všech hlavních formulářů.*

Příklad 2.1–16: *Na stránkách píšou, že to tak je* ***kvůli stovkám optimalizacím*** *kódu...*

Příklad 2.1–17: *...jsou podle něj definovány* ***tisíci každodenními kontakty****...(PDT)*

S pomocí seznamu kvantifikačních substantiv s genitivní valencí lze tento problém snadno odhalit.

Závěr

Z této kapitoly implementujeme řešení většiny problémů: obě zvrhnutí substantiv, zvrhnutí genitivního adjektiva v lokálové a několik konkrétních případů kvantifikátorů, po nichž kontrolujeme pád. Zvrhnutí lokálového adjektiva v genitivní ani špatné skloňování dvou substantiv za sebou není obecně automaticky odhalitelné.

2.2 Syntaktická a významová redundance

Redundance je dalším tématem, o kterém dosud není k dispozici žádný ucelený materiál. Podstatou je nadbytečné užití zejména spojek a adverbií přinášejících informaci, která již je ve větě obsažena. Většinou jde o nevhodnou – redundantní – kombinaci dvou slov významově i jazykově podobných.

2.2.1 Dvě ekvivalentní slova těsně za sebou

Příklady s čistými spojkami

Příklad 2.2–1: *. . . ,* ***ale zato*** *pracuje s balíčky i rpm databází o poznání rychleji.*

Příklad 2.2–2: *. . . v orientaci v hluboko odsazených částech programu nám* ***však ale*** *mnoho nepomohou.*

Příklady s čistými adverbii

Příklad 2.2–3: *Smalltalk je* ***ryze čistě*** *objektově orientovaný jazyk.*

Příklady na pomezí spojek a adverbií

Příklad 2.2–4: *. . . lze na něj nahlížet* ***toliko pouze*** *jako na jakýsi „zájmový“ spolek. (Lupa)*

Příklad 2.2–5: *. . .* ***z těchto důvodů proto*** *malloc(2) alokuje paměť. . .*

Příklad 2.2–6: *Připravuje se* ***rovněž i*** *prostředí XRE. . .*

Příklad 2.2–7: *. . . bloky jsou stále adresovatelné,* ***neboli jinými slovy*** *– nebyly uvolněny.*

Případy typu 1 a 2 jsou časté, ostatní se vyskytují ojediněle. U dvou čistých spojek nebo čistých adverbií stejného významu těsně za sebou se zřejmě všichni shodnou, že jde o chybu (u spojek i syntaktickou), nicméně těchto případů je celkově menšina a mnohé pravděpodobně vznikly nepozorností autora např. při přepisování věty. U příkladů, které jsme zařadili „na pomezí“, se názory v lingvistických kruzích různí, mnozí pro ně nalézají pochopení i vysvětlení. Ačkoli stále trváme na tom, že tato spojení jsou redundantní a smysl věty by při odstranění jednoho z problémových větných členů zůstal přesně zachován (a držíme se toho při ručních opravách nám svěřených textů), nebudeme tento přísný názor vnucovat uživatelům a implementaci provádíme jen u jediného bezesporného případu z první kategorie, konkrétně u spojení *však ale*.

Toto spojení je jednoznačně chybné (význam *ale* je zahrnut ve významu *však*), v korpusu se vyskytuje často, takže implementace má smysl. Opačný výskyt (*ale však*) může mít stylistické opodstatnění (***Ale však** oni si Češi, jak je znám, brzy nastřádají na nové. (DJC)*), proto na něj neupozorňujeme.

Množinu nepřípustných dvojic sousedících slov můžeme samozřejmě kdykoli jednoduše rozšířit.

2.2.2 Redundantní spojky na hranici mezi klauzemi

Do této skupiny patří dva případy, přičemž oba vzácně splňují všechny následující charakteristiky: vyskytují se často, jsou jednoznačně špatně a dají se snadno odhalit.

Prvním z nich je kombinace *bez toho, aniž* vzniklá zřejmě kontaminací významově ekvivalentních *bez toho, aby* a *aniž*.

> **Příklad 2.2–9:** *Tím umožňuje snadné vyřazení části dokumentu **bez toho, aniž** byste museli nějak výrazně do zdrojového kódu zasahovat.*

> **Příklad 2.2–10:** *Většina bank nabízí úvěry do 100 tisíc Kč **bez toho, aniž** byste potřebovali ručitele. (Měšec)*

V obou těchto případech (i ve všech ostatních) stačí ponechat spojku *aniž* a bez náhrady zlikvidovat *bez toho*, žádné slovosledné úpravy nejsou nutné. Tuto kombinaci je snadné nalézt, neboť do řetězce *bez toho, aniž* nemůže vniknout žádné slovosledné rozvolnění, vše se koncentruje kolem čárky mezi větami. Implementace by byla možná i v obyčejném korektoru překlepů (spellcheckeru), my ji zahrnujeme do pravidel psaných v jazyce LanGR.

Podobným případem je kombinace *proto, protože*, která se vyskytuje zejména v mluveném projevu, kde je její výskyt alespoň částečně omluvitený, ovšem v projevu písemném, odkud pochází i následující příklad, ji skutečně nelze tolerovat:

> **Příklad 2.2–11:** *Ověření terminálem jsem zvolil **proto, protože** Eurotel nemá žádné heslo...*

Nalezení je opět (a z týchž důvodů) velice snadné a provádíme jeho implementaci, ovšem znění opravy nelze automaticky navrhnout, neboť záleží na obsahu a struktuře souvětí – někdy je vhodné nechat *protože* a odstranit *proto* (příklad 11), jindy je lepší kombinace *proto, že* a jsou i případy, kde není vhodný ani jeden z uvedených postupů a je třeba souvětí od základů přeformulovat.

2.2.3 Porušení struktury souvětí

V následujících příkladech tohoto oddílu je problém v tom, že spojka na začátku souvětí „předepisuje“ určitou strukturu celému souvětí, a tu pak také čtenář očekává. Tato struktura je však další spojkou porušena – je zbytečně zesílen kontrast mezi tvrzením vedlejší a hlavní klauze.

Příklad 2.2–12: ***Jelikož*** *platí, že budete muset přepisovat do assembleru jen malé části aplikace, je* ***proto*** *nejvhodnější použít inline assembler.*

Příklad 2.2–13: ***Ačkoli*** *jsou šéfredaktoři nedostupní,* ***přesto*** *si troufám odhadovat, že budou rádi spolupracovat.*

Příklad 2.2–14: ***I když*** *předposlední nastudování se zdá být osudové,* ***přesto*** *se Hello, Dolly! stala druhým nejhranějším titulem v karlínském divadle. (letáček)*

Co se týče automatického odhalení, obecně na tento problém samozřejmě se současnými prostředky nestačíme. Složitá struktura dlouhého souvětí poskytuje živnou půdu mnoha víceznačnostem, proto se zatím spokojíme se vzorovou implementací speciálních případů (kombinace konkrétních spojek v dostatečně krátkém souvětí).

2.2.4 Nevhodná rozvití a stupňování

Pro úplnost dodejme, že do této kapitoly patří i redundantní rozvíjení či stupňování různých slov, převážně adjektiv a deadjektivních adverbií. Syntakticky nepřípustnou konstrukcí je například *více problematičtější*, mezi syntakticky správné, leč významově nepřípustné patří oblíbený *hlavní protagonista*, z našich příkladů např. *mnohem definitivněji* či roztomilý nesmysl *ohrožení ztráty lidského života*, kde věta vinou redundance získává zcela jiný význam. Případům adjektiv resp. adjektivních adverbií a jejich řešení se podrobněji věnujeme v následující kapitole.

Závěr

Z úvah této kapitoly opět vyplývá implementace částečná – konkrétní nevhodné spojky za sebou, konkrétní spojky na hranici mezi větami a několik jednodušších případů porušení struktury souvětí. Většina problémů není řešitelná obecně (ekvivalentní slova za sebou, spojky na hranici mezi větami) z důvodu specifického užívání mnohých spojek, homonymie s adverbii apod.; porušení struktury souvětí není obecně řešitelné současnými prostředky, protože nemáme k dispozici spolehlivou syntaktickou analýzu.

2.3 Stupňování adjektiv a adverbií

Stupňování se týká adjektiv kvalitativních (jakostních), tedy takových, která vyjadřují vlastnost v užším smyslu (ostatní adjektiva, vyjadřující především různá určení a vztahy, se nazývají relační (vztahová) ([8], s. 178)).

Stupňování by se vždy mělo provádět užitím komparativu a superlativu, často však dochází k chybnému užití pomocných adverbií *více* a *nejvíce*, a to jak ve spojení *více* s pozitivem namísto užití komparativu (resp. *nejvíce* s pozitivem namísto užití superlativu), tak ve významově redundantních spojeních *více* s komparativem (jiný typ redundantního spojení jsme nedoložili). Zřejmě jde o vliv cizích jazyků, zejména angličtiny.

2.3.1 *více*, *nejvíce* + pozitiv

Tato spojení se při troše benevolence dají prohlásit pouze za „ošklivá“, tedy ne vyloženě gramaticky špatná, ale při kontrole stylistiky by varování samozřejmě bylo na místě.

Příklad 2.3–1: *. . . co* ***nejvíce jednoduchý****. . .*

Při vyhledávání výskytů adverbia *více* a následujícího adjektiva v pozitivu (jiným způsobem postupovat nelze, neboť syntaktickou analýzu nemáme k dispozici) bohužel způsobuje problém homonymie s číslovkou *více* resp. *nejvíce*, viz příklad 2.

Příklad 2.3–2: *. . . rodinách, kde mají* ***více dospělých*** *dcer. (ČNK)*

V části případů je následující adjektivum v množném čísle a v genitivu, tudíž je můžeme odfiltrovat. Ovšem v těch ostatních na sobě *více* a adjektivum vůbec nezávisejí, proto nám tvar adjektiva neposkytne k významu *více* žádnou informaci, viz příklad 3:

Příklad 2.3–3: ***Víc sociální*** *pracovníci dělat* ***nemohou****. . . (ČNK)*

Dalším problémem je, že existují i adjektiva nestupňovatelná (relační), u kterých stupňování pomocí přípon a předpon aplikovat nelze, dostupné automatické prostředky je však od adjektiv kvalitativních nerozliší. Z obou uvedených důvodů nemůžeme pravidlo s čistým svědomím aplikovat a od implementace ustoupíme.

2.3.2 *více* + komparativ

Tato spojení jsou jednoznačně chybná:

Příklad 2.3–4: *. . . v dnešní době, kdy se rozesilatelé nevyžádané pošty stávají* ***více a více agresivnějšími****. . .*

O homonymii *více* a záludnostech větné stavby platí totéž co u předchozí skupiny, viz příklad 5.

Příklad 2.3–5: *. . . v Evropě je* ***více menších*** *společností. . . (ČNK)*

Naštěstí zde hovoří statistika pro nás a po odfiltrování adjektiv v množném čísle (kde je značná pravděpodobnost číslovkového čtení *více*) a několika dalších speciálních případů (podrobněji v implementační kapitole) se nám podařilo nepřesnosti v užití pravidla alespoň na datech z ČNK zcela eliminovat (zbude nám kolem 60 případů skutečné chyby). Předpokládáme tedy malou pravděpodobnost chybného užití pravidla i u jiného vstupu.

2.3.3 Stupňování nepřípustné významově

Některá adjektiva (a adverbia od nich odvozená) jsou z důvodu svého „definitivního" významu nestupňovatelná a kvantitativně nerozvíjitelná – nic nemůže být *optimálnější* ani *velmi nezbytné* apod. Příklady:

Příklad 2.3–6: *. . . z první části si i neznalý čtenář udělá* ***velmi komplexní*** *obrázek o světě kolem Linuxu a OpenSource vůbec.*

Příklad 2.3–7: *. . . takže sálající horko umíme ze svých letních dnů vyškrtnout* ***mnohem definitivněji*** *než dřív. (Listy hl. m. Prahy)*

Příklad 2.3–8: *Fields jsou tedy* ***velmi nezbytné*** *v Inventoru.*

Implementaci několika konkrétních případů provádíme, obecné pravidlo je však obtížné zformulovat, protože tato slova nejsou morfologicky rozpoznatelná, je třeba mít jejich seznam.

Závěr

Z této kapitoly je implementována zhruba polovina jevů – spojení *více* a komparativu (adjektiv i adverbií) a několik konkrétních případů stupňování nepřípustného významově. Pro obecné řešení stupňování nepřípustného významově nemáme prostředky, leč seznam konkrétních slov je kdykoli jednoduše rozšiřitelný. Stupňování „nehezké" implementovat nemůžeme vzhledem k množství zcela různorodých protipříkladů a také vzhledem k nemožnosti automaticky rozpoznat kvalitativní a relační adjektiva.

2.4 Příslovečné spřežky

Tzv. *příslovečné spřežky* (viz [22], s. 52) vznikají spojením podstatných jmen, zpodstatnělých jmen přídavných, zájmen a číslovek s příslovci a předložkami. Taková spojení se pak píší dohromady a mají význam prostého příslovce.

Některé spřežky se mohou psát dohromady i zvlášť se stejným významem, jiné se píší pouze dohromady. O ty nám jde v této práci především, neboť jejich chybné psaní může vést k víceznačnosti či ke změně smyslu tvrzení.

Hranice v mnoha případech nejsou přesné, neboť ani ve výslovnosti není mezi oběma variantami výrazu rozdíl (hlavní slovní přízvuk zůstává v obou případech jediný).

Jednoslovná povaha výrazu je nejvíce znát tam, kde již vedle spřežky nemáme výraz předložkový (*stěží*, *vstříc*, *vůbec*...), tyto případy nás pro potřeby této práce nezajímají, neboť se v nich ani chybovat nemůže. Dalším případem spřežek, které není možné psát zvlášť, jsou taková spojení, jejichž význam či větná platnost se výrazně odchýlily od původního předložkového výrazu (*spatra – s patra*, *nahoru – na horu* apod.), těmi se budeme zabývat podrobněji.

V jiných případech pojetí často kolísá, protože významový rozdíl mezi spřežkou a příslovečným předložkovým výrazem není žádný nebo se projevuje jen nepatrně (*kupodivu – ku podivu*, *bezpochyby – bez pochyby* apod.). Přechod k významu příslovce je tedy často plynulý, a proto se tam, kde význam spřežky je týž jako význam původního víceslovného výrazu, připouští dvojí způsob psaní – dohromady i zvlášť. Jen dohromady se však píší spojení, která nabyla významu jednoznačně příslovečného.

2.4.1 Rozdělení spřežek podle PČP

Pravidla [22], s. 52, uvádějí toto rozdělení psaní příslovečných spřežek (zkráceno, kurzivou naše poznámky):

a) **Spojení předložek se jmény podstatnými**
   - aa) Jako jedno slovo píšeme tyto příslovečné výrazy: dohromady, dokonce, nahlas, nahoru, nakvap, nazmar, občas, vcelku, vzápětí, zčásti, zpravidla, zbrusu apod. *(Tedy takové, kde vsunutí mezery vede ke změně významu, případně vznikne nesmysl.)*
   - bb) Obojím způsobem se píšou např. tyto výrazy: bezesporu i beze sporu, bezpočtu i bez počtu, bezpochyby i bez pochyby, kupodivu i ku podivu, kupříkladu i ku příkladu... *(Je zřejmé, že zde nedochází k výrazné změně významu.)*

b) **Spojení předložek se zpodstatnělými jmény přídavnými**
   aa) Jako jedno slovo se píšou např. příslovečné spřežky: doprava, nalevo, doleva, odjakživa, potichu, zhruba, zprava, zřídka, ztěžka, zticha, zvolna aj. *(Často se chybuje v ukazatelích směru (do leva apod.), v ostatních uvedených případech jen zřídka; smysluplnost spojení po vsunutí mezery je ve většině případů diskutabilní.)*
   bb) Obojím způsobem se píšou např. tyto výrazy: donedávna i do nedávna, doširoka i do široka, nakrátko i na krátko, naprázdno i na prázdno, naživu i na živu, zblízka i z blízka, též dozelena i do zelena, namodro i na modro. . . *(Opět je zřejmé, že oba výrazy mají smysl a není mezi nimi výrazný významový rozdíl.)*

c) **Spojení předložek se zájmeny a číslovkami** se píšou většinou dohromady, např. beztoho, nadto, nato, nacož, poté, potom, proto, pročež, předtím, přesto, přičemž, přitom, vjedno, vtom, zajedno, zato; obojím způsobem se však píše např. mimoto i mimo to, poprvé i po prvé, podruhé i po druhé, pokaždé i po každé, zasvé (vzít) i za své. *(Hranice v tomto případě není příliš jasná a zřejmě se stále vyvíjí.)*

d) **Spojení předložek a příslovcí** se píšou dohromady, např. nahonem, najednou, navíc, nanejvíc, nanejvýš(e), napoprvé, napořád, napotom, naproti, napřed, napříště, navždy, nazítra, odevšad, odjinud. *(Opět jedna ze zajímavějších skupin, ve většině případů mají oba výrazy smysl, ale pokaždé jiný.)*

Teoreticky bychom měli hlídat správné psaní slov (tedy dohromady) z bodů **a/aa**, **b/aa**, **c** a **d**. Nicméně tento výčet poněkud omezíme v souladu s našimi možnostmi a také v souladu s tím, ve kterých případech se vůbec chybuje.

Není bez zajímavosti, že Česká mluvnice z roku 1986 (jde ovšem o nové vydání verze z roku 1960) [8], s. 75, ještě povoluje psaní obojím způsobem u většiny spřežek vzniklých ze substantiv, adjektiv a číslovek, psaní dohromady předepisuje pouze u spřežek vzniklých z příslovcí a z většiny zájmen. Body **a** a **b** by tedy podle ní vypadaly ještě benevolentněji.

2.4.2 Spojení předložek se substantivy

Z této skupiny (**a/aa**) nemáme (z vlastního materiálu) téměř žádný doklad špatného užití (tedy nadbytečné mezery). Navíc by odhalování bylo velmi nesnadné až nemožné, neboť většina výrazů psaných zvlášť má smysl a syntakticky se chová jako příslovce, proto nijak neporušují větnou stavbu, a nelze je tedy automaticky odhalit ani za pomoci syntaktické analýzy.

2.4.3 Spojení předložek s adjektivy

Pro část první skupiny (**b/aa**) platí totéž co pro předchozí skupinu (výrazy s mezerou mají smysl a chovají se syntakticky ekvivalentně), ovšem z druhé části se při psaní zvlášť stanou nesmysly (*do prava*, *z prava*, *na levo*, *z hruba*, *z řídka*, *z těžka*). Nicméně jediné chyby doložené z našeho materiálu se vyskytují ve špatném psaní směrových ukazatelů, viz pikantní příklad 1:

Příklad 2.4–1: *Vidíme, že proces rendrování začíná i končí u Separátoru a jeho děti jsou rendrovány* ***z leva do prava****.*

Proto se v implementaci omezíme pouze na ně (ovšem dáme pozor na případy *levo – pravý* (ČNK) apod.). Přidání dalších slov by bylo triviální, nicméně to není třeba vzhledem k tomu, že se v nich nechybuje. Samozřejmě je možné pravidla kdykoli tímto směrem rozšířit.

2.4.4 Spojení předložek se zájmeny a číslovkami

V této skupině (**c**) se chybuje s oblibou, např.

Příklad 2.4–2: *Krátce* ***na to*** *byl založen projekt Firebird…*

Příklad 2.4–3: ***Po té*** *se přepneme na záložku Link a v poli Object/library modules…*

Příklad 2.4–4: ***Po té*** *se zapíše alespoň část číselného či slovního označení objektu…*

Bohužel i v těchto případech má výraz s mezerou většinou smysl a automaticky nelze jednoznačně rozpoznat, který z výrazů do věty správně patří, např. *po té* by bylo správně, pokud by odkazovalo k nějaké akci (ženského rodu) provedené dříve, tedy zmíněné v předešlých větách (ovšem její existence nutně neznamená, že se odkazuje zrovna k ní).

Měli jsme několik hypotéz o tom, které speciální případy zachytit lze (*přes to* v pozici těsně za čárkou, *za to* tamtéž, *za jedno* za tvarem slovesa *být*), bohužel ČNK vše vyvrátil rozumnými, byť ojedinělými protipříklady.

Jediný příklad, který (až na dobře definovatelné výjimky, viz dále) vyvrácen naštěstí nebyl, je chybné psaní *zato* místo spojení *za to* vlivem homonymie se spojkou *zato*. Vyhledáváme výskyty psaní dohromady na pozici před čárkou, kde se spojka vyskytovat nesmí (příklady 5 a 6).

Příklad 2.4–5: *Měl jsem* ***zato,*** *že… (ČNK)*

Příklad 2.4–6: *Zbožňovali jsme ho* ***zato,*** *že… (ČNK)*

Výjimka nastává pouze tehdy, pokud se *zato* vyskytuje za jinou čárkou, začátkem věty, případně za další spojkou (typicky *ale*), viz příklady 7 a 8. Tehdy čárka následující pravděpodobně odděluje klauzi vloženou, za níž bude klauze obsahující *zato* pokračovat. *Zato* tedy není na konci klauze, nýbrž na začátku, a jeho výskyt tudíž není chybný.

Příklad 2.4–7: ***Zato,*** *jak dokládá dobové svědectví,. . . (ČNK)*

Příklad 2.4–8: *. . . ale* ***zato,*** *pane, ta postava! (ČNK)*

Pravidlo *zato* před čárkou (s příslušnou výjimkou) tedy implementujeme, ostatní chybná užití mezer jsou většinou neodhalitelná, případně psaní zvlášť nedává smysl, a tudíž se v něm nechybuje (*pročež* vs. *pro čež*).

2.4.5 Spojení předložek a adverbií

V těchto případech, zmíněných v odstavci **d**, se také chybuje poměrně často:

Příklad 2.4–9: *Takto jednoduše napsaný server není schopen obsluhovat více klientů* ***na jednou****.*

Příklad 2.4–10: *. . . ostatně,* ***do dnes*** *existuje spousta jejich emulátorů i na PC.*

Většina případů má při odděleném psaní také smysl, tyto významy lze ovšem částečně syntakticky rozlišit – spojení předložky a adverbia patří jen do levého adjektivního rozvití substantiva (*na jednou napsaný server*, *do dnes vyřízené žádosti. . .*) (viz též příslušnou kapitolu na straně 27). Bohužel ČNK poskytuje bohatý výběr rozvolnění a protipříkladů (způsobených hlavně tím, že mezi adverbia patří i různá „vycpávková“ slova povahy spíše částicové), vše podtrženo všudypřítomnou homonymií různých slovních druhů (na správnou morfologickou disambiguaci nemůžeme úplně spoléhat). Proto se zatím omezujeme na vyhledávání předložky a adverbia na konci věty, kde máme sice také mnoho protipříkladů (*na jindy* apod.), můžeme je však vyloučit seznamem.

Závěr

Částečná implementace výsledků této kapitoly zahrnuje obecné pravidlo pro spojení předložek a adverbií na konci věty a několik konkrétních případů spojení předložek se zájmeny a adjektivy. Z ostatních případů některé jsou nesnadno odhalitelné (u spojení předložek se substantivy víceznačnost nerozpoznatelná ani na úrovni povrchové syntaxe), pro jiné nelze zformulovat dostatečně obecné pravidlo (spojení předložek s adjektivy), ve zbylých se nechybuje.

2.5 Rozlišení významu spojky *nebo*

Spojka *nebo* může mít význam vylučovací, nebo slučovací, v případě významu vylučovacího se před ni klade čárka, v případě slučovacího nikoli. Nicméně pisatelé zde velmi často chybují; většinou v tom, že čárku nenapíší tam, kde být má, nikoli naopak.

Cíl vždy automaticky rozhodnout, zda je význam *nebo* slučovací, nebo vylučovací, je příliš ambiciózní a stěží dosažitelný (sémantika), můžeme se však pokusit odhadnout alespoň některé případy, a to významu vylučovacího. Co se týče slučovacího významu, na žádné alespoň trochu zobecnitelné jednoznačné případy jsme zatím nenarazili, většinou je možné počítat alespoň teoreticky s oběma možnostmi (uvažujeme automatické zpracování, tedy bez zapojení znalosti kontextu a reálného světa).

Při odhalování významu spojky *nebo* samozřejmě nejde o pouhé umístění čárky, zjištěná informace může přijít k užitku i později při sémantické analýze věty. Také proto jsme si z interpunkce vybrali právě tuto otázku.

2.5.1 Kritéria pro odhalení významu vylučovacího

Podezřelé konstrukce

Vedle známé konstrukce *buď(to) . . . (a)nebo*, která se uvádí jako typický vylučovací význam, jsme vypozorovali i další: *ať již/už . . . nebo* a *zda . . . nebo* (viz příklady chyb 1, 2, chybějící čárky uvedeny v závorkách). Je však třeba dát pozor na to, zda se v souvětí nevyskytuje více *nebo* (příp. *či*), tehdy patří čárka před alespoň jedno z nich, ale typicky nejsme schopni automaticky určit, před které.

> **Příklad 2.5–1:** *Ať již budeme akcionáře hodnotit ratingem**(,) nebo** prostým selským rozumem, dojdeme nejspíš v obou případech k závěru . . . (Měšec)*

> **Příklad 2.5–2:** *Zatím není jisté, zda tento fakt má**(,) nebo** nemá nějaký efekt v praxi.*

Vyhledávky v ČNK hovoří pro to, že konstrukce *buď . . . nebo* značí vylučovací význam vždy, konstrukce *zda . . . nebo* a *ať již . . . nebo* ve většině případů, pokusíme se z nich tedy specifikovat alespoň část jistých. Patří mezi ně takové, kde po spojce *nebo* následuje před koncem klauze či celé věty již jen jedno slovo, u konstrukce *zda . . . nebo* je to typicky *ne*.

U konstrukce *ať již . . . nebo* se často vyskytují případy vícečetných koordinací, které bohužel ztrácíme tím, že nedovolíme, aby se v hledaném řetězci vyskytla čárka (omezujeme se totiž na jednu klauzi, abychom předešli chybě, a jinak její hranici rozpoznat neumíme).

Slova významově opačná

Další indicií může být, jsou-li větné členy spojené pomocí *nebo* vzájemně významově opačné; v obecném případě bychom k tomuto zjištění potřebovali slovník antonym, ovšem ani ten by mnohé situace zřejmě nezachytil (příklad 1, *rating* vs. *prostý selský rozum*), v implementaci bychom se mohli omezit na případy zjistitelné morfologickou analýzou, tedy dvojice sloves, adjektiv či adverbií se stejným lemmatem, z nichž jedno je negováno a druhé nikoli.

Bohužel tato druhá podmínka není postačující, viz dvojici protipříkladů 3 a 4 – oba obsahují stejnou dvojici vzájemně významově opačných sloves a v prvním z nich je správně užito *nebo* vylučovací, kdežto ve druhém *nebo* slučovací.

> **Příklad 2.5–3:** *Třetí vlastností, kterou lze* ***vypnout, nebo zapnout****, je antialiasing bodů, který se používá pro rozmazání hran.*

> **Příklad 2.5–4:** *Antialiasing nelze* ***zapnout nebo vypnout*** *uvnitř „závorkových" příkazů glBegin() a glEnd().*

(Pozorný čtenář si po přečtení následující kapitoly o negacích jistě všimne, že v příkladu 4 by místo *nebo* mělo být *ani*, nicméně můžete si místo slovesa *nelze* představit *lze*, pak už je užití slučovacího *nebo* zcela korektní.)

Z uvedených důvodů (většinu případů nenalezneme a u nalezených si nemůžeme být jisti významem) toto kritérium neaplikujeme.

Závěr

Jednoznačné rozlišení významu každého *nebo* je (zatím) nemožné, podařilo se nám však zformulovat alespoň některá kritéria pro odhalení *nebo* vylučovacího (a vyvrátit jiná). Tato pravidla implementujeme omezeně (uvnitř konstrukce se nesmí vyskytnout čárka), přičemž hlídáme, zda se ve větě nevyskytuje ještě další kandidát na vylučovací spojku.

2.6 Zájmena a spojky v negativní klauzi

Česká mluvnice [8], s. 327, uvádí, že ve větě záporné se zájmena neurčitá nahrazují zápornými a slučovací spojovací výrazy záporkou *ani*. Častým jevem je nekorektní užití slučovacích spojek a neučitých zájmen v těchto větách. Zřejmě jde o vliv jazyků, ve kterých se kvantifikátorů a spojek užívá „logicky" (zejm. angličtiny). Otázkou je, jak se s tím vyrovná přirozený vývoj jazyka, zda konstrukce, které jejich autorům v této kapitole vytýkáme, budou časem považovány za správné, či nikoli.

Nejhorším nešvarem, na který narážíme v celé této práci, je víceznačnost; bohužel ani v této kapitole jí nebudeme ušetřeni. Konstrukce, které jsou chybné, totiž často mohou být interpretovány i jinak, s odlišných významem, leč syntakticky i stylisticky správně. Mezi těmito významy bohužel nelze rozumně rozhodnout bez znalosti kontextu, často i čtenář je na pochybách. Ale o tom podrobněji dále.

2.6.1 Kvantifikátory

Častým pochybením bývá užití kvantifikátoru *jakýkoli* v negované části věty (typicky přímo za slovesem), kde je na místě kvantifikátor *žádný*, a související chybné užití tvaru zájmena *cokoli* místo správného *nic*.

Příklad 2.6–1: *. . . již* ***nemá jakoukoli*** *cenu. . .*

Příklad 2.6–2: *. . .* ***neumožňovala jakékoliv*** *jiné použití než na demonstrační účely (Karel Kulhavý, obhajoba diplomové práce)*

Příklad 2.6–3: *. . .* ***nezískali jsme jakékoliv*** *informace. . .*

Příklad 2.6–4: *Doposud jsme v rámci seriálu používali sokety pro přenos dat způsobem, který po nás* ***nevyžadoval cokoliv*** *vědět o principech počítačových sítí.*

V příkladech 1–3 je vhodné nahradit tvar zájmena *jakýkoliv* odpovídajícím tvarem zájmena *žádný*, v příkladu 4 na místo *cokoliv* patří *nic*, i když v tomto případě se nedá tvar *cokoliv* prohlásit za vysloveně špatný.

Nicméně podívejme se na hypotetický protipříklad:

Příklad 2.6–5: *Já* ***nemám jakékoli*** *chutě (nýbrž zcela konkrétní. . .)*

V tomto případě je spojení negativního slovesa a zájmena *jakýkoli* zcela v pořádku, věta má samozřejmě zcela jiný význam. Obdobné protipříklady lze nalézt i pro zájmeno *cokoli*.

Otázkou je, zda a jak lze tyto významy automaticky rozlišit. Pomohou nám samozřejmě i dílčí úspěchy, z příkladu 2 lze např. vypozorovat, že spojení *jakékoli jiné* může potvrdit náš dojem, že jsme nalezli případ, který chceme opravit. Naopak „zdůvodnění“, které nutně musí následovat po korektním užití tvaru zájmena *jakýkoli* po negativním slovesu (v příkladu 5 je jím ono *nýbrž zcela konkrétní*), nám v hledání chybných případů nikterak nepomůže, neboť může mít jakoukoli formu a může se nalézat až v následující větě nebo ještě dále, proto na jeho neexistenci ve větě rozebírané nelze stavět.

Nicméně při hledání v ČNK (spojení negativní sloveso + tvar zájmena *jakýkoli* resp. *jakýkoliv*) jsme nenalezli žádný výskyt protipříkladu, všech cca 150 nalezených výskytů byly správně odhalené chyby. Proto se domníváme, že vzhledem k velice nízké pravděpodobnosti špatného označení můžeme v tomto případě uživatele varovat. U zájmena *cokoli* již situace tak jednoznačná není, z 28 výskytů byly dva správné a v některých případech je navíc potřeba změnit slovosled. Dále tvar zájmena *cokoli*, přestože je užit špatně, nepůsobí tak rušivě, proto by pisateli varování mohlo vadit, tudíž budeme benevolentní a raději od něj upustíme.

Při rozvolnění podmínek (1–3 libovolná slova mezi negativním slovesem a tvarem zájmena *jakýkoli*) se však již protipříklady objevují (cca 5 %), a protože je nelze bez porozumění významu textu rozpoznat, varování nevypisujeme.

2.6.2 Spojky

Dalším případem je chybné užití slučovacích spojek a částic *i*, *a*, *nebo*, *také* na místech, kam patří spojka *ani* (tedy opět v negativní klauzi), a to jak tam, kde má spojka skutečnou koordinační funkci, tak i na místech, kde má (zejména *i*, *také*) funkci pouze zdůrazňovací, částicovou. Jev také postihuje různé větné členy.

Spojka nebo částice *i*

> **Příklad 2.6–6:** *Stane-li se, že* ***i čisté zprávy neprocházejí*** *přes poštovní server… (částice, podmět)*

> **Příklad 2.6–7:** *…* ***nenaleznete*** *v něm žádné skiny nebo, chcete-li, témata* ***i další funkce****, které jsou spjaty spíše s prostředím než s GUI. (spojka, předmět)*

V obou těchto případech je vhodné nahradit *i* správnějším *ani*. Už na pohled je zřejmé, že *i* nesprávně užité ve funkci částicové není tolik na závadu jako ve funkci koordinační, kde značně ztěžuje porozumění textu.

Pro funkci částicovou existují navíc protipříklady, kde je užití *i* v negativní klauzi na místě:

Příklad 2.6–8: ***Nedám*** *v sázku* ***i Camelot****! (První rytíř)*

Příklad 2.6–9: *Každý, kdo si modul CVSROOT checkoutne, by asi* ***neměl*** *dostat* ***i soubor passwd****. . .*

Problém je v tom, že vhodnost užití *i* či *ani* poznáme pouze při porozumění významu. V příkladu 8 jde o to, zda král již něco v sázku dal a nechce Camelot přidávat (*i*), nebo ještě nic v sázku nedal (*ani*). V příkladu 9 záleží na tom, zda vývojář již něco dostal a soubor passwd by neměl potenciálně dostat k tomu (*i*), či nedostal nic (*ani*).

V tomto případě samozřejmě chybu automaticky rozpoznat nelze. Bohužel je tomu tak i v případě potenciální koordinace – při vyhledávání dvou substantiv shodujících se ve všech morfologických kategoriích a spojených pomocí *i* ještě nemusí jít o koordinaci společně závislou na předcházejím negativním slovesu. ČNK poskytuje řadu protipříkladů, kdy spolu tato substantiva nemají nic společného a každé syntakticky patří k jiné části věty.

Spojka *a*

Příklad 2.6–10: *Tuto službu* ***neposkytuje*** *Český Telecom, Eurotel, Oskar* ***a T-Mobile****. (web televize Nova) (podmět)*

Příklad 2.6–11: *Stejně tak buffer s přijatými daty* ***nebude obsahovat*** *záhlaví* ***a zápatí*** *linkového rámce. (předmět)*

Zde je situace zdánlivě snazší, neboť *a* má v drtivé většině případů funkci koordinační, ovšem krom obdoby problému předchozí spojky (dva větné členy spojené pomocí *a* zdaleka nemusejí znamenat koordinaci) se zde objevuje i dodatečný odstín významový. Spojku *a* je vhodné ponechat nejen v případě různých názvů a frazeologických spojení, nýbrž i tehdy, pokud nechceme koordinované členy od sebe významově „odtrhnout“, tzn. chceme, aby bylo zřejmé, že negativní sloveso se vztahuje k nim společně jako k nedělitelné skupině, a nikoli, že „jen“ vylučuje každý z nich zvlášť, viz příklad 12. Rozlišování těchto případů od ostatních samozřejmě automatické metody nezvládnou.

Příklad 2.6–12: *. . . abychom* ***neléčili nenávist a zlobu*** *nenávistí a zlobou. . .*

Spojka *nebo* ve slučovací funkci

Příklad 2.6–13: *Proto také* ***neobsahuje žádné funkce*** *pro práci s okny, pro vytváření grafického uživatelského rozhraní* ***nebo pro zpracování*** *událostí.*

Příklad 2.6–14: *Antialiasing* ***nelze zapnout nebo vypnout*** *uvnitř „závorkových" příkazů glBegin() a glEnd().*

Situace je obdobná jako u spojky *a* (tedy neřešitelná), navíc přibývá problém, jak rozpoznat slučovací *nebo* – na správné kladení (resp. v tomto případě nekladení) čárky nelze u pisatelů spoléhat a náš korektor (viz předchozí kapitolu) je bohužel schopen rozpoznat jen některé významy vylučovací, ale slučovací nikoli; bez porozumění textu to mnohdy ani není možné.

Částice *také*

U této částice je třeba rozlišit, stojí-li samostatně, nebo zdůrazňuje-li význam nějaké spojky (*a*, *ale*). Ve druhém případě se podle situace spojkou *ani* nahrazuje buď pouze samotné *také*, nebo celá dvojice.

Příklad 2.6–15: ***Nedošlo také*** *k vraždění židovského obyvatelstva v ghettu (Obrazový opravník obecně oblíbených omylů)*

Příklad 2.6–16: ***Nemůžete*** *nic zkazit, ale* ***také*** *příliš mnoho získat.*

V obou těchto případech je vhodné *také* nahradit spojkou *ani*. Bohužel částice (někdy též příslovce) *také* je velice živá, mnohoznačná, může stát ve větě téměř kdekoli v různých syntaktických funkcích, proto je těžké automaticky rozhodnout, kdy je zrovna v pozici nevhodné. Dobrým protipříkladem je 13 – zde záleží na kontextu, zda je vhodnější *také*, či *ani* (buď se dříve popsal nějaký důvod, který něco způsobuje a krom toho neumožňuje práci s okny (*také*), nebo se dříve hovořilo o něčem, co (nevyjádřený podmět) neobsahuje a práci s okny k tomu přidáváme (*ani*), každopádně pro správné rozhodnutí by bylo třeba určit základ a ohnisko výpovědi).

Je zřejmé, že rozhodování je zde podobně komplikované jako u částicového *i*, proto od implementace rovněž upustíme.

Závěr

Z této kapitoly je implementována jen malá část – některé dílčí případy kvantifikátorů. U spojek se nám staví do cesty zejména rozlišení jejich významu (*i*, *také*) a neschopnost jednoznačného automatického rozpoznání koordinace (všechny). Dá se předpokládat, že se syntaktickou analýzou by se mnohé zjednodušilo, ale vše bychom stejně nevyřešili.

2.7 Levé adjektivní rozvití substantiva

aneb Kabáty na ramínku

. . . že se tam k Ebíkovi příslušná kolegyně zpíjí do němoty.

Martin Pergel, 2003

Podnadpis kapitoly byl zvolen na počest oblíbeného příkladu prof. Panevové:

Příklad 2.7–1: *Dívka rovná* ***na ramínku vystavený kabát****. [15]*

Levé adjektivní rozvití substantiva je ve většině případů stylisticky nevhodné a navíc je zdrojem následujících dvou možných problémů:

1. vznik falešné větné dvojice (viz [19], s. 61–62)
2. kumulace dvou předložek vedle sebe (viz [12])

(k oběma problémům viz též [21], s. 25–30.)

Příklad 1 patří do první skupiny, zlepšit jej lze přeformulováním na *Dívka rovná kabát vystavený na ramínku*. Do této skupiny patří i příklad 2, druhou skupinu reprezentuje příklad 3, u obou pomůže větu přeformulovat a rozvití přesunout za substantivum.

Příklad 2.7–2: *. . . v několika kostelech jsem viděl* ***ze stropu visící modely*** *lodí. (Island, ostrov zrozený z ohně, dále jen Island)*

Příklad 2.7–3: *. . . v osmém týdnu dokázali zobrazovat okna* ***v do C přeloženém interpretu****. . .*

2.7.1 Výsledný slovosled po úpravě

Odložíme nyní otázku, které konstrukce tohoto typu lze vůbec automaticky nalézt, a podíváme se na možnosti automatického návrhu opravy. Ty jsou bohužel (bez syntaktické analýzy) velmi omezené, neboť ačkoli v obecném případě se rozvití přesouvá přímo za rozvíjené substantivum, někdy je toto substantivum rozvito ještě zprava. Naše rozvití se pak přesouvá až za toto pravé rozvití. Jeho rozsah však není v našich silách automaticky rozpoznat, viz ukázkový příklad 4 z webové stránky neznámého motorkáře, kdy rozvití *slibně se vyvíjející* dokonce není kam přesunout (všimněte si i dalšího nevhodného rozvití ve druhé polovině věty):

Příklad 2.7–4: *Po necelém kilometru chytám několik řádně vypasených masařek a taktéž ukončuji* ***slibně se vyvíjející kariéru*** *čmeláka slalomisty, který se po zdárném a bravurním průletu kolem*

brutálně smykem jedoucí škody Felicie zřejmě dostatečně nevěnoval řízení letu a mou přilbu zaregistroval příliš pozdě.

Pro úplnost uveďme, že v tomto příkladu problém falešné větné dvojice (*ukončuji slibně*) nevzniká, neboť *se* na Wackernagelově pozici (viz dále) jednoznačně přiřazuje *slibně* k rozvití za sebou. Tím si ale ten pěkný příklad nebudeme kazit.

Dalším problémem je slovosled přesunutého rozvití. Jde-li o typický případ předložkové skupiny rozvíjející adjektivum, bude za substantivem (a jeho případným pravým rozvitím) většinou následovat nejprve adjektivum a poté předložková skupina, ovšem najdou se i výjimky, kde lépe působí opačné pořadí. V následujícím příkladu jsou možná obě řešení a záleží na aktuálním členění.

Příklad 2.7–5: *Mezi ně patří* ***na ostrově oblíbené jaterníky****. (Island)*

Pokud bychom zdůrazňovali, že jaterníky jsou oblíbené na ostrově (a na pevnině nikoli), byl by na místě slovosled *jaterníky oblíbené na ostrově*, ovšem vzhledem k tomu, že o ostrově je celá kniha, a tudíž v této větě zmínka o něm nepřináší žádnou novou informaci, je vhodnější slovosled *jaterníky na ostrově oblíbené*.

Následuje několik příkladů ještě pikantnějších.

Příklad 2.7–6: *Na rozdíl* ***od u nás dříve běžně rozšířených značek*** *motocyklů. . . (instrukce ÚAMK)*

Příklad 2.7–7: *Než jsem se nadál, měl jsem* ***v dece zabalené jehně v ruce*** *a ona ohřívala mléko. (Island)*

Při formulování vhodné úpravy příkladu 6 se zapotí i zdatný stylista a v příkladu 7 zase úprava výjimečně zahrnuje i přesunutí rozvití *v ruce* (které rozvíjí sloveso) nalevo od substantiva (*měl jsem v ruce jehně zabalené v dece*). Tyto příklady snad již každého přesvědčí, že úpravu slovosledu skutečně nelze navrhnout automaticky.

2.7.2 Možnosti automatického nalezení

Automatické odhalení hranice levého adjektivního rozvití substantiva je bez syntaktické analýzy téměř nemožné, snadno jsme schopni rozpoznat pouze situace, kdy nevhodné rozvití způsobilo kumulaci dvou předložek. Ani zde však automaticky neodhalíme konec příslušného větného úseku (předložky mohou mít stejnou rekci, může se vyskytnout elipsa apod.). Je třeba dát pozor také na to, že dvě předložky vedle sebe mohou znamenat i pouhý překlep, a zohlednit tento fakt při formulaci varování.

Dalším speciálním případem je, když se rozvití vyskytuje na začátku věty a zprava je ohraničeno klitikou, která obvykle zaujímá

Wackernagelovu pozici, tedy druhou ve větě. Nicméně při vyhledávání těchto podezřelých sekvencí dostáváme převážně případy, kdy předložková skupina rozvíjí spolu s adjektivem až substantivum, nikoli ono adjektivum. Tyto případy nelze s pouhou morfologickou analýzou rozlišit (*na jaře příštího roku* vs. *na hlavu postaveného*...).

Ostatní výskyty zkoumaného jevu není možné přesně rozpoznat a ohraničit, zejména vinou falešných větných dvojic.

Závěr

Vzhledem k tomu, že závadnost zkoumaného jevu je v obecném případě diskutabilní a slovosledné úpravy jsou individuální, nelze automaticky zformulovat návod k opravě, ani kdybychom tento jev nalézt uměli. Současnými prostředky lze automaticky odhalit pouze případy, které kumulují dvě předložky vedle sebe (čímž také prokazují svou závadnost), proto v těchto případech vypisujeme varování a doporučujeme (nikterak konkrétně) větu přeformulovat. U dvou předložek vedle sebe dáváme pozor na to, aby bylo předložkové čtení jednoznačné, a odfiltrujeme také některé předložky nevlastní, inspirováni prací [17].

2.8 Kontaminace

Česká mluvnice [8] uvádí tuto definici: „Kontaminace (směšování vazeb) záleží v tom, že se místo náležité vazby užije, zpravidla pod vlivem vazby výrazu (slovesa) významově blízkého, vazba nová. Někdy se tyto nové vazby vžijí a stanou se variantními prostředky spisovnými. Často jsou však takové vazby projevem neznalosti nebo rozpaků, a jsou tedy nesprávné."

Z hlediska formálního jde o porušení valenčního rámce (slovesa, řidčeji substantiva) a nahrazení některé z jeho pozic nežádoucím doplněním.

2.8.1 Případy jednoznačné – teoreticky řešitelné

První skupinu tvoří takové jevy, které by s pomocí syntaktické analýzy a valenčního slovníku [20] byly snadno odhalitelné, neboť slovesa (resp. substantiva), která byla chybně doplněna, mají jednoznačný (přesněji nezaměnitelný) valenční rámec, a chybí-li doplnění obligatorní, je nesprávné užití už velice pravděpodobné. Jsou to například tyto jevy:

Valence slovesa

V některých z těchto případů si autor pravděpodobně ani neuvědomuje, že sloveso použil špatně (zhoubný vliv médií apod.), v jiných zřejmě pouze nedopatřením zapomněl původní vazbu a nahradil ji jinou s podobným významem (tzn. při opětovném přečtení by si svého pochybení všiml).

- **vyvarovat se něčemu** (vyvarovat se něčeho × vyhnout se něčemu) – obligatorní genitivní valence nahrazena dativem
- **nastat k něčemu** (nastat Nom. × dojít k něčemu, *Řízení se vrací na místo, kde **k přerušení nastalo**.*) – obligatorní nominativní valence nahrazena předložkovou skupinou *k* + dativ
- **zabývat se něčemu** (zabývat se něčím × věnovat se něčemu, ***Otázce** komprimace videozáznamu **se** patrně ještě **budeme zabývat** později.*) – obligatorní instrumentální valence nahrazena dativem
- **brát v úvahu na něco** (brát v úvahu Acc. × hledět na Acc., ***Bere v úvahu** dokonce **na vzdálenost** míst komunikace.* (Lupa)) – obligatorní akuzativní valence nahrazena předložovou skupinou *na* + akuzativ

Valence substantiva

Zde máme z vlastního materiálu jen jeden příklad:

- **průzkum do něčeho** (průzkum něčeho × průnik/pohled do něčeho, *Proto se dnes, kromě dalšího **průzkumu do nitra** balíčkovacího*

systému, podíváme...) – obligatorní genitivní valence nahrazena předložkovou skupinou *do* + genitiv

Ačkoli jsou tyto případy jednoznačné, není jejich automatické odhalení bez syntaktické analýzy nikterak snadné – to, že se „blízko" slovesa vyskytuje „podezřelý" větný člen nevyhovující valenčnímu rámci slovesa a zároveň chybí obligatorní doplnění, ještě neznamená, že jde o chybu – obligatorní doplnění může být nevyjádřené a nevhodné doplnění možná vůbec ke slovesu nepatří. Spolehlivá syntaktická analýza by tento problém vyřešila.

2.8.2 Případy víceznačné – neřešitelné

Druhou skupinu tvoří slovesa, která mají více valenčních rámců, přičemž pokud je jeden z nich použit nesprávně místo jiného, nelze toto pochybení odhalit ani s pomocí syntaktické analýzy, neboť je třeba porozumět kontextu. Jsou to např. tyto případy:

- **zabrat Acc.** × **zabrat na Acc.** (*Popsat veškeré techniky potvrzování u okna by zabralo na mnoho článků...*) – Zde je valenční rámec s obligatorním akuzativem nahrazen valenčním rámcem s předložkovou skupinou *na* + akuzativ, nicméně my automaticky nepoznáme, zda je použití korektní (*zabrat na návnadu* apod.), či nikoli.
- **stát se z něčeho něco** × **stát se někdo (z něčeho) něčím** (*Mám dojem, že se z GPdf stává zajímavým a funkčním prohlížečem...*) – V tomto případě nepoznáme, zda jde o (chybný) valenční rámec s obligatorním nominativním a *z* + genitivním doplněním, nebo o rámec s obligatorním doplněním nominativním (které není vyjádřeno přímo ve větě), instrumentálním a fakultativní předložkovou vazbou *z* + genitiv. Někdy to nepozná bez znalosti kontextu ani čtenář, automatické metody již zcela selhávají. Náš příklad je trochu přitažený za vlasy, nicméně *GPdf* může být i označení jakéhosi vývojového stadia prohlížeče.

(Srov. [19], kap. II.2.2, Homonymie vyjádření valenčního a volného doplnění.)

2.8.3 Výjimky z předchozího – řešitelné

Zajímavý je následující případ – sloveso má sice dva různé valenční rámce, leč v kontaminaci není správně použit ani jeden z nich, proto tento případ teoreticky odhalit lze:

- **těšit se něčeho** (těšit se něčemu × těšit se z něčeho, *Technologie svobodného software* ***se těší*** *čím dál* ***lepšího přijetí*** *ze strany dodavatelů.*) – Místo obligatorního dativního doplnění či alespoň alternativní

předložkové skupiny *z* + genitiv (věta má při užití obou těchto rámců přibližně stejný význam) se zde vyskytuje pouze genitiv.

Podobný příklad máme i u valence substantiva:

- **vztah mezi něčím a něčeho** (vztah mezi něčím a něčím x vztah něčeho a něčeho, ***Vztah mezi zadáním polohy*** *a orientace světelného zdroje* ***a prováděných transformací***.) – Smíšen rámec s předložkou *mezi* doplněnou dvěma instrumentály koordinovanými spojkou *a* a rámec se dvěma genitivy opět koordinovanými spojkou *a*, s valenčním slovníkem substantiv snadno odhalitelné, neboť výsledek nevyhovuje žádnému rámci – dochází zde k narušení vztahu vzájemné reciprocity (viz [16]) obou doplnění.

Závěr

Na základě této kapitoly nevzniká žádná implementace vzhledem k nutnosti syntaktické analýzy. Pokud bychom tuto analýzu měli, bylo by možné za pomoci dat z valečního slovníku implementovat část případů, nicméně s víceznačnými případy bychom si ani tehdy neporadili. V každém případě nelze porušení valence kontrolovat obecně vzhledem k přílišné volnosti jazyka (většina doplnění je fakultativních, i obligatorní mohou být nevyjádřená apod.), je třeba postupovat případ od případu.

2.9 Nadbytečnost částice *tak*

> *Kdyby chtěl pánbůh potrestat lidstvo, tak by paní Doležalová měla dvojčata.*
>
> *Jéňa Hendrych, 1997*

Slůvko *tak* ve funkci částice (obvykle v pozici těsně za čárkou) je téměř vždy nadbytečné a je vhodné jej odstranit (spolu s případným přeformulováním následující klauze). Typickým výskytem je souvětí začínající spojkou *pokud*, ve většině z nich se částice *tak* vyskytuje na začátku druhé klauze:

Příklad 2.9–1: *Pokud vám qmailadmin nebude ověřovat hesla,* ***tak*** *se přesvědčte, že…*

Příklad 2.9–2: *Pokud použijete některý example,* ***tak*** *se připravte na to, že dost souborů…*

Příklad 2.9–3: *Pokud např. zavoláte funkci, která se bude snažit modifikovat překrytou globální proměnnou,* ***tak*** *se jí to podaří.*

Příklad 2.9–4: *Pokud uvedete konstantu,* ***tak*** *všechno, co je za touto konstantou, Perl ignoruje.*

Řidčeji se vyskytují souvětí uvozená jiným způsobem, např.:

Příklad 2.9–5: *Aby se nám adresy v adresáři nehromadily,* ***tak*** *zajistíme jejich pravidelné odstraňování.*

Příklad 2.9–6: *Jak jsem později zjistil,* ***tak*** *Tripwire může být velmi užitečný u počítače, kde má práva administrátora více lidí.*

Příklad 2.9–7: *Myslím si ale, že kdo chce,* ***tak*** *tyto informace stejně získá.*

Příklad 2.9–8: *Jestli si věříte, že jste si dobře vědomi toho, co je potřeba hlídat,* ***tak*** *je lepší si vytvořit vlastní seznam souborů.*

V příkladech 1–3 a 5–8 je vhodné *tak* odstranit, na první pozici za čárkou přesunout nejbližší sloveso a zbývající slovosled dotčené klauze ponechat nezměněn. Bohužel tuto slovoslednou úpravu nelze doporučit obecně, v příkladu 4, kdy je dotčená klauze rozdělena další vloženou klauzí na dvě části, je třeba slovosled přeformulovat drastičtěji (*Pokud uvedete konstantu, ignoruje Perl všechno, co je za touto konstantou.*), stejně tak v mnoha příkladech nalezených v ČNK, kde je výsledný slovosled zcela individuální.

Vyskytují se i případy, kdy *tak* ani v částicové funkci nevadí (*když už, tak už*), jsou však ve výrazné menšině a jedná se především

o různá úsloví. Většinu z těchto konstrukcí by bylo možné vyloučit vhodně formulovanou podmínkou (*když* v úvodu předcházející klauze apod.).

Závažnějším problémem je bohužel homonymie s adverbiem *tak*, které se také může dostat do pozice za čárkou. Konstrukci *jak . . . , tak . . .* sice můžeme snadno vyloučit, ostatní případy (*. . . , tak jsem se ostatně dostal i sem*) však nikoli – přes mizivý výskyt v ČNK je nutno s nimi počítat. Dalším problémovým případem je, když se adverbium *tak* v platnosti měrové vyskytne za čárkou po větě vložené, viz příklad 11:

Příklad 2.9–11: *Byl jsem vším, co jsem viděl,* ***tak*** *překvapen. . . (ČNK)*

Závěr

Všechny tyto výjimky nás nutí od implementace ustoupit, protože se snažíme nevarovat uživatele nadbytečně, což by v tomto případě hrozilo velmi často, byť třeba pokaždé z jiného důvodu.

2.10 Koordinace v rámci předložkové skupiny

V této kapitole se zaměříme na speciální množinu předložkových skupin (Pg), u nichž se místo řídícího substantiva vyskytuje koordinace více substantiv. Zdůvodníme si, proč je v těchto případech záhodno předložku opakovat a proč z týchž důvodů nelze příslušné případy automaticky odhalit.

Markéta Straňáková [19] ve své přelomové práci o Pg uvádí tuto definici: „Termínem předložková skupina míníme předložkový pád spolu se členy na něm nepřímo zavislými, tedy ekvivalent toho, co je v bezprostředně složkových gramatikách označováno jako předložková fráze. (Protože tento pojem je úzce spjat s formalismem bezprostředně složkovým, a navíc je v češtině termín fráze terminologicky obsazen, pokládáme za vhodné používat jiný termín.) V nejjednodušším případě jde o pouhou předložku a substantivum, jehož pád je dán rekcí předložky.“

Následuje-li za předložkou rozsáhlá koordinace, je vhodné předložku přiměřeně opakovat, zvláště u vzdálenějších členů koordinace, kde hrozí riziko vytvoření falešné větné dvojice a tím chybné interpretace věty. Příklady (předložka uvedená v závorce ve větě původně nebyla a my doporučujeme ji buď přidat, nebo o tom alespoň zauvažovat (ma místech označených otazníkem)):

> **Příklad 2.10–1:** *Naše snažení tedy směřuje k dokonalé integraci jednotlivých aplikací,* ***(k?)*** *dobré podpoře národních jazyků,* ***(k?)*** *přehledné struktuře UI a dialogů a určitě také* ***(k)*** *malé velikosti a* ***(k)*** *rychlosti aplikací.*

> **Příklad 2.10–2:** *Ohledně psaní přes elektronickou poštu autor nabádá k mazání řetězových mailů,* ***(k?)*** *posílání malých souborů,* ***(k?)*** *používání hlaviček s kontakty a* ***(k)*** *trpělivosti při čekání na odpověď. (Lupa)*

> **Příklad 2.10–3:** *. . . což představuje široké možnosti použití, například při tvorbě těles pomocí CSG nebo* ***(při)*** *zobrazování stínů.*

Ve všech třech příkladech hrozí při absenci předložky bez otazníku špatná interpretace (minimálně automatická, v příkladu třetím možná i čtenářova): v příkladu 1 může jít nikoli o *směřování k malé velikosti*, nýbrž o *strukturu malé velikosti*, v příkladu 2 vzniká místo *nabádání k trpělivosti* falešná větná dvojice *používání trpělivosti* a v příkladu 3 hrozí místo správného *použití při zobrazování stínů* chybná interpretace *tvorbě těles pomocí zobrazování stínů*. Bohužel tyto případy nelze automaticky rozlišit od situace, kdy by naopak přidání předložky větu poškodilo, např. v následujícím příkladu 4:

Příklad 2.10–4: *Peníze pak používáte k zaplacení dalších staveb, ale i provozu jednotlivých linek.*

Zde skutečně jde o *zaplacení provozu* a nikoli o *použití k provozu*, ačkoli v tomto případě jsou oba významy sémanticky ekvivalentní.

Ve všech případech je problém s homonymií valenčního doplnění – oba potenciálně rozvíjené větné členy mají sice odlišnou valenci, ale v příkladech 1, 2 a 4 (dativ a genitiv) i v příkladu 3 (lokál a genitiv) jsou u homonymního koordinovaného členu možná obě tato čtení. Příklad 1 má navíc ještě jeden problém – bez opakování předložky *k* před *rychlosti* to vypadá, že by rychlost měla být také malá.

Opakování předložek, zejména neslabičných, je vhodné i na místech, kde víceznačnost nehrozí, v uvedených příkladech např. u všech ostatních koordinovaných členů. Zajímavý je i následující příklad 5, kde sice větě bez opakování předložky rozumět je, a to jednoznačně, ale působí velmi kostrbatě.

Příklad 2.10–5: *Postupem času také došlo ke značnému rozšíření knihovny OpenGL, která se kromě UNIXových grafických stanic začala používat i na počítačích řady PC s operačními systémy Windows NT a posléze i Windows 95, Windows 98 a dále* ***(s)*** *celou řadou operačních systémů založených na novém jádře NT verze 5.0.*

Na druhou stranu čím bližší koordinovaný člen, tím může přidaná předložka působit rušivěji (příklad 1 by s pěti *k* vypadal přespříliš úderně hlavně při čtení nahlas), a naopak, čím je tento člen vzdálenější, tím je těžší se ve větné struktuře automaticky zorientovat (příklad 5, kde situaci navíc komplikují nesklonná Windows).

Závěr

Na základě této kapitoly nevzniká žádná implementace, protože problém vzhledem k mnoha víceznačnostem přesahuje možnosti současných technických prostředků. Některé dílčí případy by sice automaticky řešitelné byly, jsou to však bohužel právě ty, které „tolik nevadí“ – závažnost absence předložky vyniká zejména při vzniku víceznačnosti, která nám však znemožňuje jev odhalit, čímž se dostáváme do začarovaného kruhu.

2.11 Vokalizace předložek

Vokalizace předložek byla teoreticky rozebrána Vladimírem Petkevičem a Hanou Skoumalovou v práci [18] (vokalizace předložky *s/se* byla později upřesněna Karlem Olivou v práci [14] zabývající se morfologickou disambiguací slovního tvaru *se*). V rámci své práce jsme nalezli některé chyby a nepřesnosti rozboru původního, což (mimo další podněty) vedlo k jeho úpravám a rozšířením (Vladimír Petkevič, nepublikováno). Z těchto výsledků vycházíme při implementaci, některé otázky však stále zbývá dořešit (zejména vokalizaci předložky *s/se* v souvislosti s homonymií a disambiguací *se*).

Většina vokalizačních pravidel má povahu čistě fonetickou – kontroluje, zda slovo následující po předložce patří do některého ze seznamů, případně zda skupina písmen na jeho začátku splňuje určité podmínky. Tuto kontrolu jsme implementovali ve formě perlovského skriptu, který pro zadanou předložku a následující slovo určí, zda a jak má být předložka vokalizována, podrobnosti následují v příslušné implementační kapitole na straně 69.

Složitější pravidla kladou podmínky na slovní druh, na výslovnost apod., tudíž v našem perlovském skriptu zahrnuta nejsou.

Implementací vokalizačních pravidel v jazyce LanGR se zabývá Vladimír Petkevič v rámci projektu pravidlové disambiguace (kde tato pravidla pomáhají rozhodnout v případě homonymie – *beze*, *přede*, *se* apod.).

Kapitola 3

Technické prostředky

Základním programem, bez něhož by se implementace neobešla, je morfologická analýza Jana Hajiče [7], která na základě slovníku určí u každého známého slovního tvaru lemma a různé morfologické charakteristiky (viz [5]). Většina slov má však více různých čtení, proto je vhodné aplikovat některou z metod morfologické disambiguace.

Obvykle používaná statistická disambiguace [7] vykazuje určitou chybovost, proto jsme se při našem „opatrném“ přístupu rozhodli vyhnout se jí. Zatím nedokončená disambiguace řízená pravidly [13] by měla mít chybovost minimální, ovšem za cenu toho, že některé tagy (u kterých nedokáže rozhodnout) zůstanou stále víceznačné. K možnosti kombinace obou přístupů viz např. [6].

Shrnutím budiž to, že náš korektor nemůže počítat se zcela disambiguovanou morfologií, měl by být schopen pracovat i na vstupu vůbec nedisambiguovaném, ovšem „čím disambiguovanější, tím lepší“, tedy čím více tagů je jednoznačných, tím více našich pravidel má možnost se uplatnit.

Některé problémy by byly řešitelné (nebo lépe řešitelné) za pomoci syntaktické analýzy, pro češtinu ji však dosud (v dostatečné spolehlivosti) nemáme k dispozici. Jistě by šlo celý průběh stylistické kontroly nasimulovat nad PDT (Prague Dependency Treebank, [2]), nejsme však příznivci stavění vzdušných zámků. Proto byla pravidla vyvíjena s ohledem na to, že informace o stavbě věty nemáme.

Další otázkou bylo, zda napsat vlastní program zpracovávající text a využívající na vhodných místech morfologickou analýzu, nebo použít cizí nástroj. V případě vlastního programu se též nabízely různé možnosti, jak k programování přistupovat, nejdůležitější bylo rozhodnout se mezi procedurálním a neprocedurálním programováním.

Po napsání menšího množství vlastního (procedurálního) kódu došlo nakonec k navázání spolupráce s projektem pravidlové disambiguace [13], v jehož rámci Pavel Květoň již několik let vyvíjí jazyk LanGR [11] – formalismus a programové vybavení pro psaní lingvistických pravidel (zejména disambiguačních, později i gramatických). Tento jazyk zcela vyhovoval našim potřebám, proto jsme se rozhodli stylistická pravidla implementovat v něm. To také umožnilo

vzájemnou spolupráci. Stylistická pravidla se stávají součástí gramatického korektoru a zároveň využívají jeho lingvistické identifikátory. Na základě vzájemné spolupráce také dochází k vývoji a postupné implementaci některých nových disambiguačních pravidel.

Implementační část této práce je tedy tvořena převážně pravidly psanými v jazyce LanGR a popsanými v následující kapitole.

Náš vlastní kód (napsaný převážně v Perlu) obstarává přípravu vstupu – zejména rozdělení textu na věty (s využitím kódu z podobného programu Miroslava Spousty [4] a seznamu zkratek z českého slovníku pro program ispell [9] Petra Koláře) a různé konverze. Rozdělení na věty lze použít také samostatně pomocí perlovského skriptu *vety.pl*. Dalším vlastním a přímo spustitelným kódem je vokalizace předložek (taktéž v Perlu), o níž najdete podrobnosti v příslušné kapitole na straně 69. Všechna pravidla i uvedené skripty jsou uvolněny pod licencí GPLv2, nacházejí se na CD a lze je stáhnout i z CVS projektu NLTools [4].

Kapitola 4

Popis pravidel

V této části jsou popsána všechna pravidla, která jsme implementovali v jazyce LanGR.

Vzhledem k tomu, že pravidla jsou součástí většího projektu (viz předchozí kapitolu), nelze je použít samostatně, pouze ve spojení jednak s nutným programovým vybavením (kompilátor jazyka LanGR Pavla Květoně pro převod pravidel do C++ kódu [11], morfologie Jana Hajiče [7] a konverzní skripty Pavla Květoně pro přípravu vstupu), jednak s globálními identifikátory, které byly v jazyce LanGR definovány (většinou Vladimírem Petkevičem) pro účely pravidlové disambiguace a kterých také využíváme.

Jak již bylo uvedeno v předchozí kapitole, přijímají pravidla bez újmy na obecnosti morfologicky nedisambiguovaný vstup, pro nějž jsou také optimalizována. To se projevuje např. tak, že ačkoli se snažíme držet přesnost (precision) na nejvyšší možné úrovni, občas úmyslně neodfiltrujeme některé protipříklady kvůli homonymii, protože bychom s nimi odfiltrovali téměř všechno ostatní (to se týká např. komparativních adjektiv ženského rodu v singuláru, která mají ve všech pádech stejný tvar). Jednopísmenné předložky v některých pravidlech rozpoznáváme podle formy, a nikoli podle tagu, protože každá jednopísmenná předložka dostává do vínku i druhý tag – označení příslušného písmene. Bez disambiguace bychom tedy neměli žádnou jednopísmennou předložku jistou a vstup by nikdy nemohl vyhovět podmínce v pravidle.

Pravidla dokonce obsahují ústupky případné chybné disambiguaci – kde to pravidlu neubere na přesnosti, uvolňujeme podmínky natolik, že vyhoví i homonymní rozvolnění, které by na příslušném místě vůbec nemělo co dělat. Na některých místech zase suplujeme činnost morfologie (resp. taggeru), v její současné verzi totiž existuje jen jeden tag pro slůvko *více*, ačkoli toto slůvko může mít funkci adverbia i číslovky. Protože našemu pravidlu vyhovuje pouze adverbiální čtení, je třeba význam alespoň částečně rozlišit v lokálním kontextu a číslovková čtení vyloučit.

V případě, kdy není jasné, jak by mělo vypadat lemma (*ačkoli* vs. *ačkoliv* apod.), zahrnujeme obě varianty pro případ budoucí změny morfologie.

4.1 Atrakce

4.1.1 Substantivum ovlivněné levým rozvitím

Zvrhnutí genitivu v lokál

Pravidlo vychází z faktu, že je-li substantivum v lokálu, musí se někde před ním (neoddělena slovesem) vyskytnout lokálová předložka, navíc je (pro zamezení chybného užití vlivem např. špatné disambiguace) pravidlo optimalizováno pro obvyklé výskyty této atrakce, tedy pro substantivum typu stavení v plurálu.

První varianta zachytí případy, kdy se před lokálovým substantivem a předcházejícím adjektivním rozvitím (skládajícím se jen z čehokoli skloňovatelného s možným lokálovým čtením, adverbií, spojek a částic) vyskytuje sloveso nebo nelokálová předložka.

Druhá varianta zachytí případy, kdy je před zkoumaným substantivem a rozvitím popsaným výše už jen začátek věty (libovolné slovo, které není lokálovou předložkou). Pravidlo by jistě šlo napsat obecněji (sekvence libovolných slov, která nemohou být lokálovou předložkou, následovaná lokálovým substantivem značí chybu), ovšem pak bychom si už nemohli být jisti tím, že jde o zkoumaný typ atrakce, a varování by muselo být formulováno výrazně volněji. Navíc pravidla tohoto typu již v jazyce LanGR implementoval Vladimír Petkevič.

Poslední varianta ošetřuje případ, kdy se „zvrhávající“ adjektivum vyskytuje hned na začátku věty a „zvrhlé“ substantivum následuje za ním.

Pravidlo v současné podobě bohužel nezachytí typ „na ramínku vystavený kabát“, kdy je součástí adjektivního rozvití substantiva předložková skupina (špatné *u na procesor náročných hardwarových řešeních* vs. správné *o na procesor náročných hardwarových řešeních*), rozšíření v budoucnu je samozřejmě možné.

Vzhledem k tomu, že jde jen o tento konkrétní případ, který máme dobře prozkoumaný, můžeme si dovolit odvahu (či naopak opatrnost) při kladení požadavků na adjektivní rozvití – kvůli možné chybné morfologické disambiguaci přijme pravidlo v rámci rozvolnění i adjektiva mající pouze genitivní čtení (v dobré víře, že před disambiguací měla i lokálové).

Pro úplnost uveďme, že při rozpoznávání substantiv typu stavení vycházíme z definice Vladimíra Petkeviče, která však zatím neobsahuje všechny varianty.

```
1  charset "unix";
2  parse report;
3
4  /* Atrakce, kdy substantivum, které má být genitivní, je ovlivněno
víceznačným
5     adjektivem a stane se lokálovým.
6     "u hardwarových řešeních", "přidání řady užitečných rozšířeních"
7  */
8
9  rule AtrGenLocPl {
10   rulevariant v1 {
11   /* konfliktní slovo (sloveso, nelokálová předložka/cokoli  */
12
13     ITEM true;
14     predl = ITEM ((IsSafe Verb) or ((IsSafe Preposition) and
(MustNotBe Locative)));
15     vycpavka = SEQUENCE OF ((MustNotBe (Preposition or Verb))
and (Possible
16                            (((Locative or Genitive) and Plural)
or Conjunction
17                            or Adverb or Particle)));
18     adje = ITEM ((IsSafe (Adjective and Plural)) and (Possible
(Genitive or
19                 Locative)));
20     subs = ITEM IsSafe (NounVerbal and Locative and Plural);
21
22     DIRECT REPORT "Absence lokálové předložky vylučuje lokálový
tvar substantiva
23                   " emphasize(wordformonposition(subs)) ".";
24   };
25
26   rulevariant v2 {
27   /* není vůbec žádná předložka (tedy ani lokálová) */
28
29     ITEM SentenceStart;
30     SEQUENCE of IsSafe Punctuation;
31     predl = ITEM MustNotBe (Preposition and Locative);
32     vycpavka = SEQUENCE OF ((MustNotBe (Preposition or Verb))
and (Possible
33                            (((Locative or Genitive) and Plural)
or Conjunction
34                            or Adverb or Particle)));
```

```
35      adje = ITEM ((IsSafe (Adjective and Plural)) and (Possible
 (Genitive or
36                  Locative)));
37      subs = ITEM IsSafe (Noun and Locative and Plural);
38
39      DIRECT REPORT "Absence lokálové předložky vylučuje lokálový
 tvar substantiva
40                    " emphasize(wordformonposition(subs)) ".";
41    };
42
43    rulevariant v3 {
44    /* speciální verze pro začátek věty */
45
46      ITEM SentenceStart;
47      SEQUENCE of IsSafe Punctuation;
48      adje = ITEM ((IsSafe (Adjective and Plural)) and (Possible
 (Genitive or
49                  Locative)));
50      subs = ITEM IsSafe (Noun and Locative and Plural);
51
52      DIRECT REPORT "Absence lokálové předložky vylučuje lokálový
 tvar substantiva
53                    " emphasize(wordformonposition(subs)) ".";
54    };
55  };
```

Zvrhnutí lokálu v instrumentál

Toto pravidlo má jen jednu variantu, ve které vyhledává lokálovou předložku následovanou přípustným rozvolněním (tedy sekvencí čehokoli s možným lokálovým čtením, adverbií, spojek a částic) a substantivem typu stavení v instrumentálu. Stejně jako v předchozím pravidle povolíme v adjektivním rozvití i instrumentální čtení (coby ústupek případné chybné disambiguaci).

```
1  charset "unix";
2  parse report;
3
4  /* Atrakce, kde substantivum, ktere má být lokálové, je ovlivněno
víceznačným
5     pádem adjektiva a stane se instrumentálním.
6     "při větším zatížením..."
7  */
8
```

```
 9  rule AtrLocInsSg {
10    rulevariant v1 {
11      /* lokálová předložka není uspokojena */
12
13      predl = ITEM ((IsSafe (Preposition and Not Instrumental))
 and (Possible
14                     Locative));
15      /* protože skoro každá předložka je s něčím homonymní; bezpečnější
 je
16         IsSafe Prep + IsSafe Loc, ale to se nechytí skoro nikde
 :(
17       */
18
19      vycpavka = SEQUENCE OF ((Possible (((Locative or Instrumental)
 and Singular)
20                               or Conjunction or Adverb or Particle))
 and (MustNotBe
21                               (Noun or Pronoun or Preposition or
 Verb)));
22
23      subs = ITEM ((IsSafe NounVerbal) and (IsSafe ((Instrumental
 and Singular)
24                    or (Dative and Plural))));
25      /* ústupek nedisambiguovanému vstupu */
26
27      DIRECT REPORT "Lokálová předložka " emphasize(wordformonposition(predl))
 "
28                     je ve sporu s instrumentálovým substantivem
29                     " emphasize(wordformonposition(subs)) ".";
30    };
31  };
```

4.1.2 Adjektivum ovlivněné rozvíjeným substantivem

Zvrhnutí genitivu v lokál

Pravidlo se velmi podobá pravidlu pro zvrhnutí genitivu v lokál u substantiva. Opět tu vycházíme z neexistence lokálové předložky, která je u našeho zkoumaného adjektiva obligatorní.

V první variantě hledáme konfigurace, na jejichž začátku se nachází nějaký „oddělovač“ (nelokálová předložka či sloveso), následuje přiměřené rozvolnění (viz předchozí pravidla), poté víceznačné substantivum a lokálové adjektivum.

Ve druhé variantě poslouží jako „oddělovač“ začátek věty, ve třetí variantě bez rozvolnění taktéž.

```
charset "unix";
parse report;

/* Atrakce, kdy adjektivum, které má být genitivní, je ovlivněno víceznačným
    substantivem a stane se lokálovým.
    "z modulu obsaženém", "u našeho občasníku věnovaném" ...
 */

rule AtrGenLocAdj {
  rulevariant v1 {
   /* konfliktní předložka nebo něco jiného (co nás nutně odděluje od potenciální
      lokálové předložky, která je obligatorní)
    */

     ITEM true; /* aby nasledujici nebyla prvni ve vete */
     predl = ITEM ((IsSafe Verb) or ((IsSafe Preposition) and (MustNotBe Locative)));
     vycpavka = SEQUENCE OF ((MustNotBe (Preposition or Verb)) and (Possible
                ((((Locative or Genitive) and Singular) or Conjunction or Adverb
                or Particle)));
     subs = ITEM ((IsSafe (Noun and MasculineInanimate and Singular)) and
                 (Possible (Genitive or Locative)));
     /* spíš pro formu, aby bylo vidět, z čeho se to zvrhlo */

     adje = ITEM ((IsSafe (Adjective and Locative and Singular)) and (Possible
                 MasculineInanimate));
     /* Possible kvůli homonymii s Neutrem */

     DIRECT REPORT "Absence lokálové předložky vylučuje lokálový tvar adjektiva
                   " emphasize(wordformonposition(adje)) ".";

   };

```

```
33    rulevariant v2 {
34    /* od začátku věty není lokálová předložka */
35
36      ITEM SentenceStart;
37      SEQUENCE of IsSafe Punctuation;
38      predl = ITEM MustNotBe (Preposition and Locative);
39      vycpavka = SEQUENCE OF ((MustNotBe (Preposition or Verb))
 and (Possible
40                  (((Locative or Genitive) and Singular) or Conjunction
 or Adverb
41                  or Particle)));
42      subs = ITEM ((IsSafe (Noun and MasculineInanimate and Singular))
 and
43                   (Possible (Genitive or Locative)));
44      adje = ITEM ((IsSafe (Adjective and Singular and Locative))
 and (Possible
45                   MasculineInanimate));
46
47      DIRECT REPORT "Absence lokálové předložky vylučuje lokálový
 tvar adjektiva
48                     " emphasize(wordformonposition(adje)) ".";
49    };
50
51    rulevariant v3 {
52    /* na začátku věty jen substantivum a adjektivum */
53
54      ITEM SentenceStart;
55      SEQUENCE of IsSafe Punctuation;
56      subs = ITEM ((IsSafe (Noun and MasculineInanimate and Singular))
 and
57                   (Possible (Genitive or Locative)));
58      adje = ITEM ((IsSafe (Adjective and Singular and Locative))
 and (Possible
59                   MasculineInanimate));
60
61      DIRECT REPORT "Absence lokálové předložky vylučuje lokálový
 tvar adjektiva
62                     " emphasize(wordformonposition(adje)) ".";
63    };
64  };
```

4.1.3 Špatné skloňování po kvantifikátoru

Jednoduché pravidlo vychází ze (zatím krátkého) seznamu množstevních údajů, které mají genitivní valenci (*řada*, *většina* apod.). Následuje-li po některém z nich (odděleno případným rozvolněním složeným z adjektiv, adverbií a částic) substantivum bez možného genitivního čtení, vypisuje pravidlo varování.

```
1  charset "unix";
2  parse report;
3
4  /* Atrakce, kdy substantivum za množstevním údajem přebírá jeho
pád místo toho,
5     aby naplnilo genitivní valenci tohoto údaje - "v řadě případech"
6   */
7
8  rule AtrMnoz {
9    rulevariant v1 {
10
11      predl = ITEM Possible (AnyPOS);
12      mnoz = ITEM ((Possible lemma == "řada") or (IsSafe lemma
== "většina")
13                   or (IsSafe ((lemma == "stovka") or (lemma ==
"desítka") or
14                   (lemma == "tisíc"))));
15      /* hack kvůli přechodníku řadě */
16
17      vycpavka = SEQUENCE OF ((IsSafe (Adjective and Plural)) or
(IsSafe (Adverb
18                              or Particle)));
19      ceho = ITEM ((IsSafe (Noun and Plural)) and (MustNotBe Genitive));
20
21      if (((mnoz Possible Locative) and (ceho IsSafe Locative))
or
22         ((mnoz Possible Dative) and (ceho IsSafe Dative)) or
23         ((mnoz Possible Instrumental) and (ceho IsSafe Instrumental)))
then
24         { DIRECT REPORT "Množstevní údaj " emphasize(wordformonposition(mnoz))
"
25                         vyžaduje genitiv, nikoli tvar
26                         " emphasize(wordformonposition(ceho))
"."; };
27    };
```

```
28  };
```

4.2 Syntaktická a významová redundance

4.2.1 Dvě ekvivalentní slova těsně za sebou

Pravidlo pouze vyhledává jednu konkrétní dvojici slov, která by neměla následovat hned po sobě. Rozšíření do budoucna možné, v tom případě bude seznam takovýchto slov zařazen do zvláštního identifikátoru.

```
1  charset "unix";
2  parse report;
3
4  /* Dvě slova téhož významu za sebou. Možno přidávat další varianty.
*/
5
6  rule RedDveSlova {
7    rulevariant v1 {
8
9      prvni = ITEM IsSafe (lower form == "však");
10     druha = ITEM IsSafe (lower form == "ale");
11
12     DIRECT REPORT "Spojky " emphasize(wordformonposition(prvni))
" a
13                   " emphasize(wordformonposition(druha)) " za
sebou jsou
14                   reduntantní, jednu z nich vyhoďte.";
15   };
16 };
```

4.2.2 Redundantní spojky na hranici mezi klauzemi

Pravidlo implementujeme pro dva konkrétní případy „kombinovaných“ spojovacích výrazů, v každé variantě jeden. Vyhledává přesnou sekvenci forem (s čárkou mezi nimi).

```
1  charset "unix";
2  parse report;
3
4  /* Smíchání dvou spojovacích výrazů vedoucí ke vzniku redundance.
*/
5
6  rule RedDveSpojky {
7    rulevariant v1 {
```

```
 8
 9      bez1 = ITEM IsSafe (lower form == "bez");
10      toho = ITEM IsSafe (lower form == "toho");
11      carkax = ITEM IsSafe Comma;
12      aniz = ITEM IsSafe (lower form == "aniž");
13
14      DIRECT REPORT "Spojení " emphasize(wordformonposition(bez1))
 "
15                    " emphasize(wordformonposition(toho)) "
16                    " emphasize(wordformonposition(carkax)) "
17                    " emphasize(wordformonposition(aniz)) " je
redundatní,
18                    použijte pouze " emphasize(wordformonposition(aniz))
 ".";
19    };
20
21    rulevariant v2 {
22
23      proto = ITEM IsSafe (lower form == "proto");
24      carkax = ITEM IsSafe Comma;
25      protoze = ITEM IsSafe (lower form == "protože");
26
27      DIRECT REPORT "Spojení " emphasize(wordformonposition(proto))
 "
28                    " emphasize(wordformonposition(carkax)) "
29                    " emphasize(wordformonposition(protoze)) "
je redundatní.";
30    };
31  };
```

4.2.3 Porušení struktury souvětí

Pravidlo implementujeme pro dva konkrétní případy porušení struktury souvětí, a to v omezené míře (mezi dvěma redundantními spojkami se nesmí vyskytovat čárka, která by mohla indikovat další klauzi, jejíž existence by naše rozhodování ztížila). Pravidlo bylo třeba rozdělit do dvou variant, neboť v jednom případě vyhledáváme na začátku věty jedno slovo a ve druhém dvě.

```
1  charset "unix";
2  parse report;
3
4  /* Porušení struktury souvětí (valence spojek) */
```

```
 5
 6  rule RedStruct {
 7    rulevariant v1 {
 8
 9     ITEM SentenceStart;
10     SEQUENCE of IsSafe Punctuation;
11      prvni = ITEM IsSafe ((lemma == "ačkoliv") or (lemma == "ačkoli"));
12      vycpavka = SEQUENCE OF ((Possible AnyPOS) and (MustNotBe
Punctuation));
13      carkax = ITEM IsSafe Punctuation;
14      druha = ITEM IsSafe (lower form == "přesto");
15
16      DIRECT REPORT "Spojky " emphasize(wordformonposition(prvni))
" a
17                    " emphasize(wordformonposition(druha)) " jsou
redundantní,
18                    jednu z nich vyhoďte a dolaďte slovosled.";
19    };
20
21    rulevariant v2 {
22
23      ITEM SentenceStart;
24      SEQUENCE of IsSafe Punctuation;
25      prvni = ITEM IsSafe (lower form == "i");
26      poprvni = ITEM IsSafe (lemma == "když");
27      vycpavka = SEQUENCE OF ((Possible AnyPOS) and (MustNotBe
Punctuation));
28      carkax = ITEM IsSafe Punctuation;
29      druha = ITEM IsSafe (lower form == "přesto");
30
31      DIRECT REPORT "Spojky " emphasize(wordformonposition(prvni))
"
32                    " emphasize(wordformonposition(poprvni)) "
a
33                    " emphasize(wordformonposition(druha)) " jsou
redundantní,
34                    jednu z nich vyhoďte a dolaďte slovosled.";
35    };
36  };
```

4.3 Stupňování adjektiv a adverbií

4.3.1 *více* + komparativ

Pravidlo vyhledává spojení *více* resp. *méně* a komparativu, a to jen v případech, kdy tato slova stojí těsně za sebou, bez rozvolnění. První varianta se týká adjektiv a z důvodu homonymie adverbií *více* resp. *méně* s číslovkami musí vyloučit mnoho případů, ve kterých se může objevit číslovkové čtení, příklady z ČNK jsou uvedeny v komentářích.

Nelze bohužel říci, že bychom číslovkové čtení vyloučili docela, ale při testování na datech z ČNK se zdařilo. Primárně to samozřejmě není problém náš, nýbrž morfologie a morfologické disambiguace, bohužel současná verze morfologie přiřazuje *více* i *méně* pouze adverbiální čtení, tudíž tento problém neřeší.

Ve druhé variantě pravidla řešíme spojení *více* resp. *méně* s komparativem adverbia, také musíme vyloučit některé protipříkady, ale je jich podstatně méně.

```
1  charset "unix";
2  parse report;
3
4  /* Více, méně + komparativ je špatně, je ale třeba dát pozor
na homonymii
5     s číslovkou "více", proto tolik výjimek.
6   */
7
8  rule StupVicKomp {
9    rulevariant v1 {
10
11     pred = ITEM Possible (AnyPOS);
12     vice = ITEM IsSafe (((lemma == "hodně") or (lemma == "málo"))
13                         and Comparative);
14     adject = ITEM IsSafe (Adjective and Comparative);
15     po = ITEM IsSafe Singular;
16     /* bohužel to musíme chytat až tady, IsSafe singulár u adjektiva
se nechytí
17        skoro nikde, samá homonymie
18      */
19
20     if adject Possible (Dative and Singular) then { fail; };
21     /* přejí více slabšímu */
22
```

```
23      if adject Possible (Genitive and MasculineAnimate) then {
 fail; };
24      /* více kvalitnějšího kapitálu */
25
26      /*  obecně genitiv vyhodit nemůžeme kvůli disambiguaci, protože
 by nám to
27          vyhodilo skoro všechno
28       */
29
30      if pred Possible Punctuation then { fail; };
31      /* obě mé děti, více mladší dcera */
32
33      if po Possible ConjunctionCoordinate then { fail; };
34      /* o 50% a více dražší, o nic méně či více ... */
35
36  //    DIRECT REPORT "Po " emphasize(wordformonposition(vice))
 " nesmí následovat
37  //                komparativ (" emphasize(wordformonposition(adject))
 ").";
38    };
39
40    rulevariant v2 {
41
42      vice = ITEM IsSafe (((lemma == "hodně") or (lemma == "málo"))
 and
43                          Comparative);
44      advkomp = ITEM IsSafe (Adverb and Comparative);
45
46      if advkomp Possible ((lemma == "spíše") or (lemma == "spíš")
 or
47                           (lemma == "raději")) then { fail; };
48
49      if advkomp Possible ((lower form == "méně") or (lower form
 == "samozřejmě"))
50                           then { fail; };
51
52  //    DIRECT REPORT "Po " emphasize(wordformonposition(vice))
 " nesmí následovat
53  //                  komparativ (" emphasize(wordformonposition(advkomp))
 ").";
54    };
55  };
```

4.3.2 Stupňování nepřípustné významově

Pravidlo řeší nevhodnost stupňování a jiného zdůrazňování konkrétní množiny slov, která ze své povahy stupňovatelná nejsou a zároveň se u nich pisatelé často nechávají unést. První varianta vyhledává komparativy, druhá spojení se zdůrazňovacími adverbii.

```
1  charset "unix";
2  parse report;
3
4  /* zdůrazňování adjektiv/adverbií, která ze svého významu zdůrazňovat
nepotřebují
5     "nezbytnější", "velmi optimální" ...
6  */
7
8  rule StupNemoz {
9    rulevariant v1 {
10
11     stup = ITEM IsSafe (Id_StupNemoz and Comparative);
12
13     DIRECT REPORT "Použití komparativu " emphasize(wordformonposition(stup))
 "
14                   není vhodné vzhledem k vyznamu slova.";
15   };
16
17   rulevariant v2 {
18
19     zdur = ITEM IsSafe ((lower form == "velmi") or (lower form
 == "značně"));
20     /* možno rozšířit: příliš, hodně - ale jsou homonymní */
21     stup = ITEM IsSafe Id_StupNemoz;
22
23     DIRECT REPORT "Slovo " emphasize(wordformonposition(stup))
 " není nutné
24                   zdůrazňovat slůvkem " emphasize(wordformonposition(zdur))
 ".";
25   };
26 };
```

```
1  charset "unix";
2
3  /* slova, která jsou významu "konečného", nelze je tedy stupňovat
ani jinak
4     zdůrazňovat (velmi..)
5  */
6
7  parse PositionType;
8
9  shared predex Id_StupNemoz;
10
11 Id_StupNemoz =
12      ((lemma == "definitivní")
13    or (lemma == "finální")
14    or (lemma == "optimální")
15    or ((lemma == "zbytný") and Negative)
16    or (lemma == "definitivně")
17    or (lemma == "finálně")
18    or (lemma == "optimálně")
19    or ((lemma == "zbytně") and Negative));
```

4.4 Příslovečné spřežky

4.4.1 Spojení předložek s adjektivy

Pravidlo vyhledává chyby v psaní směrových údajů a upozorňuje na to, že by se tyto údaje měly psát dohromady.

```
charset "unix";
parse report;

rule SprLevoPravo {
  rulevariant v1 {

    predl = ITEM Possible (AnyPOS);
    smer = ITEM Possible (AnyPOS);
    potom = ITEM MustNotBe Hyphen;

    private StringTuple PrepSmer;
    PrepSmer = {wordformonposition(predl), wordformonposition(smer)};

    if PrepSmer member of Id_SprLevoPravo then {

    DIRECT REPORT "Spojení " emphasize(wordformonposition(predl))
" a
                  " emphasize(wordformonposition(smer)) " se
coby příslovečná
                  spřežka píše dohromady.";
    };
  };
};
```

```
charset "unix";

/* směrové údaje, které se mají psát dohromady */

parse PositionType;

shared StringTupleList Id_SprLevoPravo;

Id_SprLevoPravo =
  [
    {"do", "prava"},
```

```
12      {"do", "leva"},
13      {"z", "prava"},
14      {"z", "leva"},
15      {"v", "pravo"},
16      {"v", "levo"},
17      {"na", "pravo"},
18      {"na", "levo"}
19    ];
20
```

4.4.2 Spojení předložek se zájmeny a číslovkami

Pravidlo vyhledává výskyt slova *zato* před čárkou, který ve většině případů vypovídá o tom, že to není spojka, nýbrž chybně napsané spojení *za to*. Vyloučíme případy, kdy před *zato* stojí čárka nebo jiná spojka, protože tehdy *zato* být spojkou může.

```
1  charset "unix";
2  parse report;
3
4  /* Homonymie spojky "zato" a výrazu "za to". "Zato" se až na
výjimky nesmí
5     vyskytovat před čárkou.
6   */
7
8  rule SprZato {
9    rulevariant v1 {
10
11     predtim = ITEM ((MustNotBe Punctuation) and (MustNotBe Conjunction));
12     /* zato není na začátku klauze ani za jinou spojkou */
13     zato = ITEM (IsSafe (lower form == "zato"));
14     carkax = ITEM IsSafe Punctuation;
15
16     DIRECT REPORT "Slůvko " emphasize(wordformonposition(zato))
" na pozici před
17                    interpunkcí nemůže mít funkci spojky, proto
se píše zvlášť.";
18   };
19 };
```

4.4.3 Spojení předložek a adverbií

Pravidlo vyhledává konfiguraci předložka – adverbium – konec věty, která typicky vypovídá o tom, že jde o chybně (s nadbytečnou mezerou) napsanou příslovečnou spřežku. Předložky nás nezajímají všechny, ale jen ty, které mohou tvořit spřežky. Je třeba vyloučit mnoho výjimek, jednak chyb morfologie, jednak konkrétních případů, kdy předložka a adverbium na konci věty stát mohou (vycházíme z ČNK, výčet pravděpodobně není úplný). Využití tohoto materiálu předpokládáme také v budoucnu při implementaci příbuzného disambiguačního pravidla.

```
1  charset "unix";
2  parse report;
```

```
 3
 4  /* spřežky typu předložka + adverbium na konci věty by se měly
psát dohromady */
 5
 6  rule SprPrepAdvEnd {
 7    rulevariant v1 {
 8
 9      predl = ITEM IsSafe Id_SprPrep;
10      adve = ITEM IsSafe Adverb;
11      SEQUENCE of IsSafe Punctuation;
12      ITEM SentenceEnd;
13
14      private StringTuple PrepAdvJ;
15      PrepAdvJ = {wordformonposition(predl), wordformonposition(adve)};
16
17      if PrepAdvJ member of Id_SprPrepOther then { fail; };
18      /* výjimky všelikého typu */
19
20      if adve Possible Dative then { fail; };
21      /* přejí více slabšímu */
22
23      if adve Possible Id_SprPrepAny then { fail; };
24      /* výjimky, které se mohou psát zvlášť s libov. předložkou
 */
25
26      if (((predl IsSafe (lower form == "do")) or (predl IsSafe
 (lower
27         form == "na"))) and (adve IsSafe Id_SprPrepDoNa)) then
 { fail; };
28      /* výjimky, které mohou stát za předložkami "do" nebo "na"
 */
29
30      DIRECT REPORT "Předložka " emphasize(wordformonposition(predl))
 " a
31                    adverbium " emphasize(wordformonposition(adve))
 " by se
32                    měly psát dohromady jako příslovečná spřežka.";
33    };
34  };
```

```
 1  charset "unix";
 2
 3  /* předložky, které tvoří spřežky */
```

```

parse PositionType;

shared predex Id_SprPrep;

Id_SprPrep = (Preposition and
    ((lower form == "do")
  or (lower form == "na")
  or (lower form == "od")
  or (lower form == "s")
  or (lower form == "v")
  or (lower form == "ve")
  or (lower form == "z")
  or (lower form == "za")
  or (lower form == "ze")));

```

```
charset "unix";

/* předložky a příslovce, která spolu netvoří spřežku */

parse PositionType;

shared StringTupleList Id_SprPrepOther;

Id_SprPrepOther =
  [
    {"do", "pryč"},
    {"do", "nikam"},
    {"v", "polosedě"},
    {"v", "pololeže"},
    {"z", "předloni"},
    {"na", "doma"},
    {"na", "ven"},
    {"pro", "doma"},
    {"pro", "ven"},

    {"na", "déle"},
    {"na", "dnes"},
    {"na", "dýl"},
    {"na", "furt"},
    {"na", "krátko"},
    {"na", "kdy"},
```

```
27      {"na", "nikdy"},
28      {"na", "pak"},
29      {"na", "potom"},
30      {"na", "teď"},
31      {"na", "těžko"},
32
33      {"od", "dřív"},
34      {"od", "dříve"},
35      {"od", "minule"},
36      {"od", "posledně"}
37    ];
38
```

```
 1  charset "unix";
 2
 3  /* příslovce i "příslovce", která netvoří spřežku s žádnou předložkou
*/
 4
 5  parse PositionType;
 6
 7  shared predex Id_SprPrepAny;
 8
 9  Id_SprPrepAny =
10      ((form == "NATO")
11    or (form == "PPP")
12    or (lower form == "id")
13    or (lower form == "atd")
14    or (lower form == "apod")
15    or (lower form == "atp")
16    or (lower form == "resp")
17    or (lemma == "málo")
18    or (lemma == "hodně")
19    or (lower form == "moc")
20    or (lower form == "tolik"));
21
```

```
 1  charset "unix";
 2
 3  /* příslovce, která netvoří spřežku s "do" a "na" */
 4
 5  parse PositionType;
 6
 7  shared predex Id_SprPrepDoNa;
```

```
 8
 9  Id_SprPrepDoNa =
10      ((lower form == "jindy")
11    or (lower form == "kdykoli")
12    or (lower form == "kdykoliv")
13    or (lower form == "později")
14    or (lower form == "pozítří")
15    or (lower form == "pozejtří")
16    or (lower form == "zítra")
17    or (lower form == "zejtra"));
18
```

4.5 Rozlišení významu spojky *nebo*

Pravidlo vyhledává některé případy chybějící čárky před spojkou *nebo*, má-li tato spojka velmi pravděpodobný vylučovací význam. Je omezeno na jednoznačné případy, tedy takové, kdy se ve větě vyskytuje jen jedno *nebo*, a my jsme si tudíž jisti, že čárka patří před něj.

První varianta hledá slůvko *buď* (homonymii se slovesem v tomto případě pomíjíme, opět ústupek nedokonalosti morofologického značkování resp. disambiguace vstupu) a kdekoli za ním *nebo* bez čárky, v těchto případech je vylučovací význam téměř jistý (v ČNK nebyl vyvrácen). Následující varianty hledají výrazy *zda* resp. *ať již* a opět kdekoli za nimi *nebo* bez čárky, v těchto případech si však natolik jisti být nemůžeme, proto se omezujeme na situace, kdy za *nebo* následuje jen jedno slovo do konce věty či klauze, tehdy jde typicky o výběr ze dvou variant.

```
1  charset "unix";
2  parse report;
3
4  /* Řešení některých případů vylučovacího nebo - poznáme a doporučíme
čárku */
5
6  rule NeboCarka {
7    rulevariant v1 {
8
9      /* "buď" a někde za ním "nebo" bez čárky, nikde jinde ve
větě není další
10        "nebo" nebo "či"
11       */
12
13     prvni = ITEM lower form == "buď";
14     vycpavka = SEQUENCE OF ((Possible AnyPOS) and (MustNotBe
Punctuation)
15 and not (lower form == "či" or lower form == "nebo"));
16     nebo = ITEM lower form == "nebo";
17     druha_vycpavka = SEQUENCE OF ((Possible AnyPOS) and not ((lower
18                 form == "nebo") or (lower form == "či")));
19     SEQUENCE of IsSafe Punctuation;
20     ITEM SentenceEnd;
21
22
```

```
23      DIRECT REPORT "Slůvo " emphasize(wordformonposition(prvni))
 " implikuje
24                  vylučovací význam spojky " emphasize(wordformonposition(nebo))
 ",
25                    je tedy třeba dát před ni čárku.";
26    };
27
28    rulevariant v2 {
29
30      /* "zda", někde za ním jediné "nebo" a pak už jen jedno slovo
 a konec věty
31          či klauze
32       */
33
34      prvni = ITEM IsSafe (lower form == "zda");
35      vycpavka = SEQUENCE OF ((Possible AnyPOS) and (MustNotBe
 (Punctuation or
36                 (lower form == "či") or (lower form == "nebo"))));
37      nebo = ITEM IsSafe (lower form == "nebo");
38      druha_vycpavka = ITEM Possible AnyPOS;
39      carkax = ITEM IsSafe Punctuation;
40
41      DIRECT REPORT "Slůvko " emphasize(wordformonposition(prvni))
 " implikuje
42                  vylučovací význam spojky " emphasize(wordformonposition(nebo))
 ",
43                    je tedy třeba dát před ni čárku.";
44    };
45
46    rulevariant v3 {
47
48      /* "ať už/již", někde za ním jediné "nebo" a pak už jen jedno
 slovo a konec
49          věty či klauze
50       */
51
52      prvni = ITEM IsSafe (lower form == "ať");
53      poprvni = ITEM (IsSafe ((lower form == "už") or (lower form
 == "již")));
54      vycpavka = SEQUENCE OF ((Possible AnyPOS) and (MustNotBe
 (Punctuation or
55                 (lower form == "či") or (lower form == "nebo"))));
```

```
56      nebo = ITEM IsSafe (lower form == "nebo");
57      druha_vycpavka = ITEM Possible AnyPOS;
58      carkax = ITEM IsSafe Punctuation;
59
60      DIRECT REPORT "Spojení " emphasize(wordformonposition(prvni))
 "
61                    " emphasize(wordformonposition(poprvni)) "
implikuje vylučovací
62                    význam spojky " emphasize(wordformonposition(nebo))
", je tedy
63                    třeba dát před ni čárku.";
64    };
65  };
```

4.6 Zájmena a spojky v negativní klauzi

4.6.1 Kvantifikátory

Jednoduché pravidlo kontroluje, zda se po negativním slovesu nevyskytuje tvar zájmena *jakýkoliv*. Rozvolnění bohužel není možné vzhledem k řadě protipříkladů, viz příslušnou teoretickou kapitolu na straně 23.

```
1  charset "unix";
2  parse report;
3
4  /* Chybné užití neurčitého zájmena (kvantifikátoru) za negativním
slovesem,
5     správně má být užito zájmeno záporné; "nemá jakoukoli cenu"
6   */
7
8  rule NegKvant {
9    rulevariant v1 {
10
11     x = ITEM IsSafe (Verb and Negative);
12     y = ITEM IsSafe ((lemma == "jakýkoliv") or (lemma == "jakýkoli"));
13
14     DIRECT REPORT "Po negativním slovesu " emphasize(wordformonposition(x))
" by
15                   místo neurčitého zájmena " emphasize(wordformonposition(y))
"
16                   měl stát odpovídající tvar zájmena žádný.";
17   };
18 };
```

4.7 Levé adjektivní rozvití substantiva

Pravidlo hledá výskyt dvou jistých předložek vedle sebe, které pravděpodobně indikují nevhodně užité levé adjektivní rozvití substantiva. Je však třeba pamatovat i na to, že se může jednat o pouhý překlep. Dále je třeba z nalezených dvojic předložek vyloučit chyby morfologie, nevlastní předložky, které se kumulovat mohou, a také nedostatky tagování – předložkový výraz o dvou slovech bývá bohužel tagován stejně jako dvě nezávislé předložky za sebou.

```
1  charset "unix";
2  parse report;
3
4  /* Dvě předložky za sebou jsou znakem nevhodného levého adjektivního
rozvití
5     substantiva, které by se mělo přesunout za ono substantivum.
6   */
7
8  rule DvePrep {
9    rulevariant v1 {
10
11     prvni = ITEM IsSafe Preposition;
12     druha = ITEM IsSafe Preposition;
13
14     if prvni Possible Id_NevlPrep then { fail; };
15     if druha Possible Id_NevlPrep then { fail; };
16     if prvni Possible Id_CiziPrep then { fail; };
17     if druha Possible Id_CiziPrep then { fail; };
18
19     DIRECT REPORT "Dvě předložky " emphasize(wordformonposition(prvni))
" a
20                   " emphasize(wordformonposition(druha)) " vedle
sebe značí buď
21                   překlep, nebo nehezké levé rozvití, které je
lepší přehodit
22                   doprava od rozvíjeného substantiva.";
23   };
24 };
```

```
1  charset "unix";
2
3  /* nevlastní předložky, které se mohou kumulovat s jinými předložkami,
a části
```

```
4     předložkových výrazů
5   */
6
7   parse PositionType;
8
9   shared predex Id_NevlPrep;
10
11  Id_NevlPrep = (Preposition and
12      ((lower form == "kromě")
13    or (lower form == "mimo")
14    or (lower form == "místo")
15    or (lower form == "namísto")
16    or (lower form == "vedle")
17    or (lower form == "podle")
18    or (lower form == "nehledě")
19    or (lower form == "vzhledem")
20    or (lower form == "pomocí")
21    or (lower form == "narozdíl")));
22
```

```
1   charset "unix";
2
3   /* slova, která Hajič taguje jako předložky, ale většinou se
tak nechovají */
4
5   parse PositionType;
6
7   shared predex Id_CiziPrep;
8
9   Id_CiziPrep = (Preposition and
10      ((lower form == "in")
11    or (lower form == "von")
12    or (lower form == "van")
13    or (lower form == "de")
14    or (lower form == "des")
15    or (lower form == "on")
16    or (lower form == "di")
17    or (lower form == "without")
18    or (lower form == "ad")));
19
```

4.8 Vokalizace předložek

Vokalizace předložek byla v omezené míře (nezahrnuje speciální případy kontroly slovních druhů, výslovnosti apod.) implementována v Perlu, příslušný skript a použité moduly naleznete na CD. Skript by se měl dát intepretovat Perlem verze 5 a vyšší a nevyžaduje nic kromě standardních perlovských modulů *Exporter* a *locale* a přiložených modulů *Vocalization.pm* a *Technics.pm*.

Skript *vocal.pl* má převážně demonstrační účel, proto vypisuje všechny typy výsledků, jaké lze z vokalizační funkce dostat, konkrétně:

- jak má vypadat vokalizace („e“, „u“ pro obligatorní vokalizaci příslušnou samohláskou, „n“ pro zakázanou vokalizaci, „j“ pro volitelnou vokalizaci)
- zda je vokalizace použitá v dotazu korektní (návratová hodnota 1 a hláška „OK“ v případě, že ano, nebo návratová hodnota 0 a podrobná rada, co s vokalizací provést, v případě, že ne)

Veškerý kód týkající se rozhodování o vokalizaci je soustředěn v modulu *Vocalization.pm*, lze jej tedy snadno volat i z jiného programu. Každá předložka zde má svou vlastní funkci, která obsahuje převážně seznamy slov a začátků slov spadajících do té či oné vokalizační kategorie.

Použití: *vocal.pl předložka následující_slovo*

Kapitola 5

Zhodnocení

Pravidla jsme testovali na PDT (1,5 milionu ručně morfologicky disambiguovaných slov, syntaktickou informaci nevyužíváme), abychom se vyhnuli chybám způsobeným nedokonalou disambiguací, které by nás zřejmě postihly při testování na statisticky disambiguovaném ČNK. Kdybychom pravidla naopak testovali na vstupu zcela nedisambiguovaném, uplatnil by se jich pouze zlomek, proto bychom nemohli zhodnotit jejich úspěšnost. Při testování na PDT se nám tak podařilo vyloučit co nejvíce cizích vlivů – morfologická disambiguace není předmětem této práce a s postupem času snad bude uspokojivě vyřešena.

Dá se předpokládat, že na vstupu „méně disambiguovaném“ či zcela nedisambiguovaném se našich pravidel uplatní méně, zároveň však můžeme vzhledem k jejich opatrné formulaci předpokládat, že počet chybných užití oproti disambiguovanému vstupu nevzroste.

Bohužel nelze dost dobře zhodnotit úplnost (recall) našich pravidel – neznáme totiž celkový počet výskytů zkoumaných jevů v testovacích datech. Není v lidských silách a mnohdy ani v možnostech techniky všechny tyto chyby nalézt. Naopak u konkrétních podmnožin zkoumaných jevů, které jsme prohlásili za odhalitelné a jejichž hledání jsme implementovali, pravidlo jednoduše nalezne vše, co vyhovuje podmínce. Tedy stručně řečeno: „nalezneme vše, co slibujeme, ale nevíme a nedokážeme zjistit, kolik z toho, co [jsme neslíbili & je chybné & nalézt by šlo], ještě zbývá“.

Celkově se pravidla při procházení PDT uplatnila 65krát, z toho 59krát dobře a šestkrát špatně. Všechna mylná užití byla způsobena chybnou ruční disambiguací, pravidla nikdy neselhala na korektně označkovaném vstupu.

Problémy byly zejména s pravidlem vyhledávajícím dvě předložky, kterého se jednak týkalo pět z šesti případů chybné disambiguace (písmeno homonymní s předložkou bylo označkováno jako předložka), jednak jsme pro něj museli zavést spoustu výjimek. Nepříjemnosti nám působily cizí předložky, ty jsme tedy z hledání vyloučili, a dále předložkové výrazy, jejichž oba členy jsou, po našem soudu neprávem, značkovány jako předložka (*nehledě na*, *vzhledem k* apod.). Také jsme z hledání vyloučili všechny tyto výrazy, na kterých

se pravidlo uplatnilo, ale je potřeba nějaké systémové řešení – předložkový výraz určitě není ekvivalentní dvěma předložkám za sebou a při vývoji nových morfologických značek by se tento problém měl zohlednit. V každém případě je nutné si dobře rozmyslet „ostré“ použití pravidla, zejména do té doby, než se morfologické značkování předložkových výrazů uspokojivě vyřeší.

Nyní k správným užitím pravidel. PDT se „bohužel“ skládá z relativně kvalitních, redigovaných textů, proto se jen málo pravidel uplatnilo více než desetkrát, některá se naopak neuplatnila vůbec.

Suverénně vedou chybějící čárky před *nebo* (31 výskytů, z toho 10 *buď*, 9 *ať již/už* a 12 *zda*), za nimi následují atrakce postihující skloňování po množstevním údaji (14 výskytů, z toho 11 *tisíc* + inst. (*se sedmi tisíci zaměstnanci*, dva *tisíc* + loc. (*na tisících stránkách*), jeden *stovka* + inst. (*stovkám dětem*)). Co se týče skloňování po množstevním údaji *tisíc*, situace se zřejmě liší podle toho, zda jde o určitý počet tisíců (první příklad *se sedmi tisíci zaměstnanci*), kdy substantivum pád číslovky předcházející *tisíc* přejmout může, nebo o prosté *tisíce* (druhý příklad *na tisících stránkách*), kde by jej přejímat nemělo, ale přesto se tak děje. V každém případě se toto skloňování stává územ a měli bychom na něj pohlížet shovívavě.

Dalším častěji se vyskytnuvším jevem, kde se naše pravidlo uplatnilo, byly dvě předložky za sebou (sedm výskytů, z toho však pětkrát byl na vině překlep a jen dvakrát hledané levé adjektivní rozvití substantiva). Ostatních jevů bylo odhaleno jen malé množství – tři výskyty neurčitého zájmena v negativní klauzi, po jednom výskytu chybného psaní spřežky na konci věty, chybného psaní *zato* před čárkou, stupňování nepřípustného významově a chybějící lokálové předložky. Neuplatnily se některé typy atrakcí, spřežek a chybného stupňování, dále se neuplatnilo žádné pravidlo pro redundanci.

Jak již bylo řečeno, k žádnému chybnému užití, které by nebylo způsobeno chybou disambiguace, nedošlo. Přesnost (precision) je tedy 100 %, problémem zůstává nízká úplnost (recall). Při opatrné formulaci pravidel jsme ani jiný výsledek očekávat nemohli, ovšem prostor pro vylepšování zde je, byť si zřejmě vyžádá hlubší analýzu zkoumaných jevů a hlavně jejich kontextu a možných rozvolnění. Nemáme bohužel k dispozici rozumně morfologicky označnovaný korpus „ošklivých“ textů, proto nemůžeme říci, zda je nízký počet uplatnění našich pravidel zaviněn jejich nedokonalostí, nebo příliš vysokou kvalitou analyzovaných dat. Protože se však většina pravidel uplatnila alespoň jednou, prokázalo se, že jsou funkční a schopná vybrané typy chyb lokalizovat.

Literatura

[1] Český národní korpus
<http://ucnk.ff.cuni.cz>

[2] Pražský závislostní korpus
<http://ufal.ms.mff.cuni.cz/pdt>

[3] Zpravodajský server Root.cz, 1999-2004, 4Web, Internet Info
<http://www.root.cz>

[4] Projekt NLTools – open source nástroje pro zpracování přirozeného jazyka
<http://sourceforge.net/projects/nltools>

[5] Hajič J.: Positional Tags: Quick Reference (Czech Morphology), 2000
<http://quest.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html
>

[6] Hajič J., Krbec P., Květoň P., Oliva K., Petkevič V.: Serial Combination of Rules and Statistics: a Case Study for Czech Tagging. In: Proceedings of the Conference of the 39th Annual Meeting of the Association for Computational Linguistics. CNRS – Institut de Recherche en Informatique de Toulouse and Université des Sciences Sociales. Toulouse, 260–267, 2001

[7] Hajič J.: Disambiguation of Rich Inflection, Karolinum – Charles University Press, Prague, 2004.

[8] Havránek B., Jedlička A.: Česká mluvnice, SPN Praha, 1986

[9] Kolář P.: Český slovník pro korektor překlepů ispell, 4. mezinárodní konference Ekonomika a informatika na přelomu tisíciletí, 13. – 14. 9. 1999, Liberec.
<ftp://ftp.vslib.cz/pub/unix/ispell/>

[10] Kučera K.: Vzpomínky na korpusového buditele, In: festšrift (k padesátým narozeninám Nikiho Petkeviče), ÚTKL, pp. 97–101, 2004

[11] Květoň P.: Rule-based Morphological Disambiguation, in prep.

[12] Machová S.: Dvě předložky vedle sebe. In: Naše řeč 1, pp. 30–34, 2000

[13] Oliva K., Hnátková M., Květoň P., Petkevič V.: The Linguistic Basis of a Rule-Based Tagger of Czech. In: Proceedings of the Text, Speech and Dialogue conference TSD 2000 held in Brno 2000. LNAI 1902, Springer–Verlag Berlin Heidelberg, pp. 3–8, 2000.

[14] Oliva K.: Linguistics-based PoS-tagging of Czech: Disambiguation of *se* as a Test Case. In: Proceedings of the 4th Conference on Formal Description of Slavic Langauges held in Potsdam, pp. 299–314, 2001

[15] Panevová J.: Funkční styly a automatické zpracování jazyka. In: Česká slavistika. České přednášky pro XII. mezinárodní sjezd slavistů, Krakov 1998, Slavia, Slovanský ústav AV ČR, pp. 161–167

[16] Panevová J.: Česká reciproční zájmena a slovesná valence. In: Slovo a slovesnost 60, pp. 268–275, 1999.

[17] Panevová J.: Existuje chyba v syntaxi? In: Sborník prací Filozoficko-přírodovědecké fakulty Slezské univerzity v Opavě, Prof. M. Jelínkovi k narozeninám, Řada D3, 2003, 145–153.

[18] Petkevič V., Skoumalová H.: Vocalization of Prepositions. In: Linguistic Problems of Czech. Final Research Report for the JRP PECO 2824 project, pp. 147–157, 1995.

[19] Straňáková M.: Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování, ÚFAL/CKL Technical report, 2001

[20] Straňáková-Lopatková M., Žabokrtský Z.: Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In: LREC-2002, Proceedings, vol. III., ELRA, pp. 949–956, 2002
<http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/>

[21] Uhlířová L.: Knížka o slovosledu, Academia, 1987

[22] Kolektiv autorů: Akademická pravidla českého pravopisu, Academia, 1998