Program pro automatickou opravu stochastické syntaktické anotace (analytická rovina PDT)

1. Základní charakteristiky opravného programu

Opravný program je počítačový software umožňující vyhledání a opravy chyb v automatické syntaktické anotaci (syntaktická anotace podle analytické roviny PDT[1]). Opravy jsou založeny na lingvistických pravidlech a seznamech vlastností slov, přičemž v současné době je implementováno 26 různých opravných pravidel. Opravná pravidla jsou koncipována tak, aby spolehlivě opravovala určitou chybnou strukturu. Program je implementován v jazyce PERL, jeho vstupem a výstupem je tzv. „vertikála“. Opravná pravidla jsou sice primárně zaměřena na výstup McDonaldova MST parseru, opravují jeho typické chyby, ale principiálně mohou korigovat výsledky jakéhokoli parseru trénovaného na PDT a implementovaného ve stejném prostředí, možná s menší účinností (parser založený na jiném principu bude dělat jiné chyby), ale výstup parseru zlepší.

1.1 Program a jeho datový zdroj

Program Fix_A_Tree_Ling_Rules.pl obsahuje cca 6000 řádků kódu, jeho datový zdroj (seznamy slov a jejich vlastností) Fix_A_Tree_Ling_Data.tsv má cca 18000 položek. Program ke spuštění vyžaduje instalované knihovny jazyka PERL (spustitelný je jak v OS Linux, tak v OS Windows). Kódování vstupu i výstupu je Unicode (UTF8).

1.2 Vstup a výstup programu

Program pracuje s tzv. „vertikálou“, na niž lze snadno převést původní PML-formát souborů PDT, popř. pracovní formát MST parseru či formát CONLL. Tato vertikála se používá jak pro vstup, tak pro výstup programu. Každé slovo ve větě je ve vertikále na jednom řádku a uvádí se pro něj šest proměnných: forma, lemma, morfologická značka, syntaktická funkce, pořadí ve větě (vyjádřené číslem) a číslo řídícího větného členu. Příklad takové vertikály následuje:

```
Výrobci   výrobce   NNMP1-----A----     Sb   1    2
deklarují deklarovat     VB-P---3P-AA---     Pred 2    0
hmotnosti hmotnost  NNFP4-----A----     Obj  3    2
výrobků   výrobek   NNIP2-----A----     Atr  4    3
v    v    RR--6----------     AuxP 5    8
nabídkových    nabídkový AAIP6----1A----     Atr  6    7
listech   list NNIP6-----A----     Adv_Co    7    5
a    a    J^-------------     Coord     8    2
na   na   RR--6----------     AuxP 9    8
```

1 http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/html/

```
dodacích  dodací    AAIP6----1A----     Atr  10   11
listech   list NNIP6-----A----     Adv_Co    11   9
.    .    Z:-------------     AuxK 12   0
```

1.3 Koncepce programu a pravidel

Program se skládá ze základní části, která analyzuje jednotlivé věty textu, a z jednotlivých opravných pravidel, která jsou samostatnými podprogramy, které hlavní program v případě potřeby spouští. Všechna pravidla jsou koncipována tak, aby bylo minimalizováno riziko nevhodného zásahu do správné struktury, pravidla jsou při opravách velmi opatrná, a pokud není (téměř) jisté, že nepoškodí správnou strukturu, raději neprovedou žádnou změnu.

1.4 Běh programu

Program ve svém průběhu nejprve načítá lingvistická data, jež využije pro ověření všech analyzovaných větných struktur. Potom postupně načítá jednotlivé věty textu s jejich závislostní strukturou, morfologickými značkami, lemmaty atd. U některých větných členů zjišťuje hodnoty dalších lingvistických charakteristik, které nejsou v původní struktuře uvedeny, aby je mohl využít při analýze věty. Tyto parametry vyplývají z kontextu a přímo z vlastností jednotlivých slov, program například rozlišuje mezi koordinačními spojkami, které tvoří hranice klauzí, a těmi, které stojí uvnitř věty.

Potom program opakovaně prochází strukturu věty, přičemž nejprve vyhledává možné chyby v základní závislostní struktuře (hlavní a vedlejší věty, závislosti sloves a spojek) a potom vyhledává ostatní pravděpodobně chybné struktury slovo po slově. Narazí-li na „podezřelou“, pravděpodobně chybnou strukturu, spustí opravné pravidlo (což je samostatný podprogram), které strukturu ověří a případně se ji pokusí opravit. Zjištění chyby je většinou snazší než její automatická oprava. Pravidlo má obvykle více možností, jak chybu opravit, pokud ale žádná z nich není v dané situaci použitelná, oprava se neprovede.

1.5 Použité seznamy slov a jejich vlastností

Opravný program využívá několik seznamů slov, v nichž jsou upřesněny některé jejich syntakticky relevantní vlastnosti. Seznamy pocházejí z rozsáhlých korpusů (SYN2005 a SYN2010, v některých případech i další korpusy Českého národního korpusu z řady SYN), a mohou tak být výrazně obsáhlejší než údaje, které parser může získat z trénovacích dat. Seznamy zahrnují valenci sloves a adjektiv, slovesa reflexiva tantum, rekci substantiv, typické adverbiální předložkové fráze, substantiva označující osoby, složené předložky, adjektiva, která mohou být syntaktickými substantivy aj. Krátké seznamy slov (jako modální a fázová slovesa) jsou zabudovány přímo v

opravném programu, nejsou načítány z externího seznamu.

Valence sloves zohledňuje reflexivitu, valence může být uvedena jak s prostým, tak s předložkovým pádem (pád je vyjádřen číslem): *zabránit 3*, *ptát se 2*, *přemýšlet o 6, pustit se do 2*, v některých případech je valence uvedena i negativně (netranzitivní slovesa). Seznam obsahuje cca 5000 údajů o valenci.

Valence adjektiv rozlišuje pouze mezi prostými a předložkovými pády: *dbalý 2*, *zodpovědný za 4*, v seznamu je cca 400 adjektiv.

U substantiv jsou v jednom seznamu uvedena jak „valenční“ (deverbativní) substantiva s valencí jako *nakládání s 7*, tak další substantiva často rozvíjená neshodným přívlastkem v určitém prostém nebo předložkovém pádu: *atentát na 4*, *spor o 4*, *hráč na 4*, *věrnost 3*. Seznam zahrnuje cca 400 substantiv.

Slovesa, která jsou označena jako reflexiva tantum, se rozdělují do dvou kategorií: patří sem jednak slovesa, která jako nereflexivní vůbec neexistují (*smát se*), jednak slovesa, která mají jako zvratná jiný význam (*hodit se*; zde označení „reflexivum tantum“ není zcela namístě), celkem je v seznamu cca 1000 sloves.

Seznam typických adverbiálních předložkových frází se využívá především pro určení syntaktické funkce v rámci předložkové fráze (***Adv*** oproti ***Obj***), někdy i pro určení závislosti. Seznam obsahuje spojení předložka – pád – lemma/forma substantiva (*o 6 přestávce*, *na 4 oplátku*, *za 4 rok*), celkem to je cca 1400 spojení s adverbiálním významem.

Substantiva označující osoby jsou zařazena do zvláštního seznamu především kvůli spojením typu *pan Novák*, *sopranistka Eva Urbanová*. V seznamu je cca 1200 apelativ (*pan, sopranistka*); 900 křestních jmen (*Jan*, *Eva*) a 2000 příjmení (*Novák, Urbanová*).

Víceslovné předložkové výrazy jako *v souvislosti s*, *na základě* nebo *směrem k* se v PDT označují jako složené předložky, opravný program proto potřebuje jejich seznam pro rozlišení, se kterými předložkovými obraty je třeba takto zacházet, a se kterými ne. V seznamu je cca 100 složených předložek.

Některá adjektiva mohou ve větě hrát roli syntaktického substantiva, aniž by bylo ve větě jasně elidované syntaktické substantivum. Část z nich by měla být disambiguována jako substantiva, např. *pracující*, *hovězí*, část je bližší zájmenům, např. *ostatní*, *další*, část si zachovává adjektivní charakter, ale často se používá bez substantiva, elidované je obecné apelativum *člověk, muž, žena*: *přítomný, umírající.* Pokud nejde o elipsu, nemají mít taková adjektiva funkci ***ExD*** (funkce slova, jehož původní řídící člen byl elidován), ale syntaktickou funkci podle vlastní funkce ve větě. Seznam zahrnuje cca 100 adjektiv s touto charakteristikou.

1.6 Určení hranic klauzí

Jedním z důležitých lingvistických parametrů, které opravný program na začátku rozboru každé věty přidává ke spojkám a interpunkčním znaménkům, je rozlišení mezi slovy, jež od sebe oddělují dvě klauze, a těmi, které stojí uvnitř věty. Mnoho chyb, kterých se parser dopouští, vyplývá právě z jeho neschopnosti pracovat s hranicemi klauzí. Systém dokáže určit, že daná spojka či interpunkce je hranicí klauzí, například když vlevo i vpravo od slova stojí sloveso v určitém tvaru. Dokáže také vyloučit hranici klauzí, např. když na koordinační spojce závisejí dvě syntaktická substantiva ve stejném pádu a spojka není jedinou potenciální hranicí klauzí mezi dvěma slovesy.

1.7 Určení sloves v určitém tvaru a reprezentantů složených slovesných tvarů

V PDT jsou věty (klauze) reprezentovány slovesy v určitém tvaru nebo plnovýznamovým slovesem ve složeném slovesném tvaru (složený minulý čas, složený budoucí čas, kondicionál, pasivum). Pro efektivní práci se základní strukturou věty je třeba tato slovesa rozpoznat a odlišit je od pomocných sloves (s funkcí ***AuxV***) či infinitivních předmětů modálních či fázových sloves (u obou typů má infinitiv funkci ***Obj***). Rozlišení nevyžaduje žádný sofistikovaný algoritmus.

1.8 Určení hlavních a vedlejších vět

Slovesa v určitém tvaru (tak budeme nadále nazývat i slovesné tvary, které jen reprezentují složený slovesný tvar, např. infinitiv u budoucího času), program označí jako řídící slovesa klauzí (reprezentují celou klauzi). Pro práci s větnou strukturou je třeba je podrobněji rozlišit na slovesa reprezentující hlavní a vedlejší věty.

Je-li na slovese v rámci klauze přímo či nepřímo závislé předcházející vztažné zájmeno, před nímž stojí přijatelný antecedent (syntaktické substantivum se stejným rodem a číslem), bude sloveso považováno za reprezentanta vztažné věty.

Pokud je sloveso v jedné klauzi s předcházející podřadicí spojkou, která stojí těsně po čárce či souřadicí spojce, bude označeno jako sloveso reprezentující vedlejší větu. Sloveso by správně mělo být závislé na podřadicí spojce, ale ne vždy parser strukturu takto skutečně označí). Algoritmus je složitější, např. spojky *jako* a *než* těsně po čárce nemusí vždy uvozovat vedlejší větu (1).

(1) *Poklidné doby akciových investorů jsou pryč , **jako** mávnutím kouzelného proutku sletěly kapitálové trhy v letních měsících o desítky procent .*

Slovesa, která nejsou označena jako slovesa ve vztažné či spojkové vedlejší větě, jsou považována za potenciální slovesa v hlavní větě. Algoritmus nedokáže spolehlivě identifikovat slovesa, která patří do vedlejší věty přerušené jinou vloženou klauzí, jako třeba sloveso

rozmazlován v příkladu (2).

(2) *Byla pevně přesvědčená o tom , že Michael , kterého tiše zbožňovala , byl* ***rozmazlován*** *v tom nejhorším slova smyslu ,*

1.9 Oprava základní struktury věty

Po získání všech potřebných informací o větě začne opravný program ověřovat základní strukturu věty složenou z formálního syntaktického kořene, spojek a sloves reprezentujících věty. Základní strukturu je nutné opravit nejdříve, protože chyba zde by mohla zablokovat opravu jednotlivých větných členů. Opravný program rozpozná několik typicky chybných struktur. Pokud na ně u dané věty narazí, spustí odpovídající opravné pravidlo.

1.10 Oprava jednotlivých větných členů

Po opravách základní struktury prochází opravný program postupně všechna slova věty. Pokud svým kontextem, závislostí, syntaktickou funkcí, morfologickou značkou nebo jejich kombinací odpovídá jedné z typických chyb větných členů, opět se spustí odpovídající opravné pravidlo, ověří, zda je struktura skutečně chybná, a případně se pokusí chybu napravit. Pokud je struktura změněna, načtou se parametry věty znovu, aby další pravidla již mohla pracovat s novou, opravenou strukturou.

1.11 Uložení věty do vertikály

Po ověření celé věty se věta zapíše do opraveného souboru ve formátu vertikály a načte se další věta.

2. Jednotlivá opravná pravidla

V následující části představíme jednotlivá implementovaná opravná pravidla, rozdělená do skupin podle svého zaměření. V souladu s cílem této práce (získat statistické údaje o syntaktických funkcích substantiv ve vztahu k jejich předložce a pádu) se největší důraz klade na opravu závislostí a syntaktických funkcí substantiv. Některá pravidla se ale zaměřují na obecnější jevy nebo na základní závislostní strukturu složených vět (souvětí), protože možnost provést dílčí opravu závislosti či funkce je často podmíněna správnou celkovou strukturou. Každá jednotlivá oprava tak může přispět k větší účinnosti pravidel specifičtějších.

Pravidla jsou implementována jako samostatné podprogramy opravného programu. Když hlavní

větev programu narazí na podezřelou strukturu, vyvolá odpovídající pravidlo (podprogram).

2.0 Způsob prezentace jednotlivých pravidel

U popisovaných pravidel vždy uvádíme nejprve stručnou motivaci, potom příklady chyb, které by pravidlo mělo opravit, podmínky, za kterých se z hlavního programu spouští, a základní postup, kterým chyby opravuje (pokud je algoritmus velmi složitý, může být výklad rozčleněn do několika částí), dále příklady řešení, počet zásahů v korpusu SYN2005 a celkovou úspěšnost pravidla.

2.0.1 Příklady chyb

Uvedené příklady chyb parseru vždy pocházejí z korpusu SYN2005 označkovaného MST parserem. Příklady oprav představují opravy provedené pomocí pravidlového opravného programu, a to popisovaným pravidlem. Zvýrazňujeme pouze chyby týkající se popisovaného pravidla stejně jako opravy (pokud jiné pravidlo opraví jinou část struktury, oprava může být v příkladu zobrazena, ale nebude nijak zvýrazněna). Příklady chyb a jejich řešení jsou uváděny ve dvou formátech podle složitosti. Složité struktury jsou představeny graficky, jednodušší textově. Opravný program často musí zpracovávat a úspěšně opravuje mnohem rozsáhlejší struktury, než zobrazujeme v příkladech, nemůžeme je však na omezeném prostoru adekvátně znázornit.

2.0.1.1 Textové zobrazení

V textovém formátu je uvedena věta nebo část věty z korpusu (bez dalších úprav, tj. zachováváme oddělenou interpunkci, malá a velká písmena, případně i textové chyby aj.). U slov, kterých se oprava týká, jsou za lomítkem uvedeny relevantní údaje: závislost, syntaktická funkce nebo část morfologické značky (obvykle prvních pět znaků pro identifikaci pádu, popř. čísla a rodu substantiva). Závislost je reprezentována číselným údajem nebo znakem pro formální kořen: závislost na formálním syntaktickém kořeni je označena dvojitým křížkem (#), závislost na jiném větném členu je vyjádřena číslem vyjadřujícím vzdálenost řídícího větného členu (záporná čísla doleva, kladná doprava).

U příkladů chyb s opravou jsou za lomítkem dva údaje, nejprve chybný, potom opravený. Závislosti, funkce a tagy jsou barevně zvýrazněny. Původně **správné údaje** jsou zobrazeny **modrou barvou**, **chyby** jsou zvýrazněny **červeně**. U příkladů oprav je **zeleně** zvýrazněna **správná oprava**, **oranžově** je zvýrazněná **oprava chybná** (tj. oprava, jejíž výsledek není zcela správný bez ohledu na správný nebo chybný vstup).

Příklad chyby:

(1) *Vyšplhá/**Pred**/**2** -/**1** li/**#** na konec druhého dílu , získává/**Adv**/**-6** 2 body .*

Sloveso *vyšplhá* má chybnou syntaktickou funkci ***Pred***, je správně závislé na spojce *li* (o dvě slova doprava). Spojovník je správně závislý na následujícím slově: spojce *li*. Spojka *li* je chybně určena jako řídící uzel celé závislostní struktury (závisí přímo na formálním kořenu věty). Sloveso *získává* má chybnou syntaktickou funkci ***Adv*** a je chybně závislé na spojce *li* (šest slov doleva).

Příklady chyby s opravou:

(1) *Vyšplhá/**Pred**/**Adv**/**2** -/**1** li/**#**/**6** na konec druhého dílu , získává/**Adv**/**Pred**/**-6**/**# 2 body .*
(2) *když jsme ji o/**6**/**2** to/**Obj**/**Adv**/**-1** souvisleji/**1** a trochu stranou požádali .*

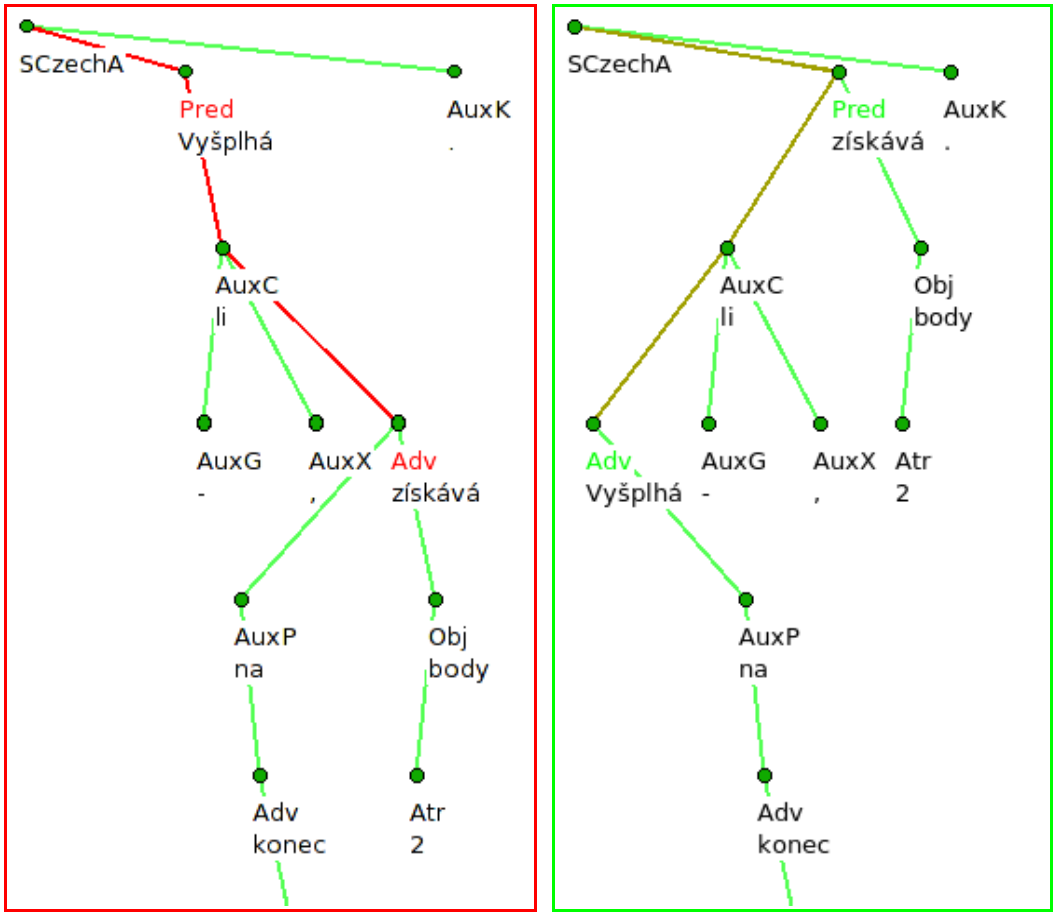
Kromě předešlého popisu chyb je **zeleně** doplněna oprava syntaktických funkcí a závislostí. V příkladu (2) došlo k chybné opravě slov *o* a *to* (oprava ze správné funkce či závislosti na špatnou), chybné opravy jsou zvýrazněny **oranžovou** barvou.

2.0.1.2 Grafické zobrazení

U složitějších případů je do textu zařazen obrázek graficky znázorněné závislostní struktury, a to celé věty nebo její části, vytvořený pomocí programu TrEd. Vždy se zobrazují dva obrázky: obrázek výchozí struktury označkované MST parserem a obrázek závislostní struktury po opravě. Výchozí struktura je červeně orámována, relevantní chyby jsou zvýrazněny červenou barvou. Chyby, na něž není dané pravidlo zaměřené, se nezvýrazňují. Opravená struktura je orámována zeleně, správně opravené závislosti a funkce jsou zvýrazněny zeleně (funkce světlezelenou barvou, závislosti žlutozelenou). Opět se nezvýrazňují chyby, které se netýkají komentovaného pravidla, ani jejich opravy. Formální syntaktický kořen věty je označen **SCzechA**. U příkladu vždy uvádíme nejprve větu v textové podobě.

Příklad grafického znázornění chybné struktury s opravou:

(1) *Vyšplhá – li na konec druhého dílu , získává 2 body .*



V uvedeném příkladu prezentované pravidlo opravilo závislosti slov *vyšplhá*, *li a získává*. Jiné pravidlo zároveň opravilo závislost předložkové fráze reprezentované předložkou *na*, tato oprava však v příkladu není zvýrazněna. Zatímco textových příkladů uvádíme pro každé pravidlo několik, a to nejprve typické chyby, potom chyby s opravami, u grafických příkladů většinou uvádíme jen jeden typický příklad s opravou.

2.0.2 Počet zásahů v korpusu SYN2005 a hodnocení úspěšnosti

Po příkladech chyb a jejich řešení uvádíme u každého pravidla počet zásahů v korpusu SYN2005, přepočítaný na 1 000 000 tokenů. Jako jeden zásah se počítá změna u jednoho tokenu. Změny rozdělujeme podle typu na 1) změny závislosti, 2) kombinované změny závislosti a syntaktické funkce (včetně případné změny morfologické značky), 3) změny syntaktické funkce (beze změny závislosti), 4) kombinované změny syntaktické funkce a morfologické značky a 5) změny morfologické značky (bez dalších změn). Pokud v průběhu zpracování věty zasáhne týž token více pravidel, počítá se jen první změna (první aplikované pravidlo).

Nakonec představujeme odhad úspěšnosti pravidla spočítaný na vzorku 100 náhodných aplikací pravidla (každý zásah pochází z jiného textu, texty pocházejí z různých funkčních stylů). Změny

hodnotíme třemi stupni: úspěšná změna, neutrální zásah, chybná změna. Za **úspěšný zásah** považujeme změny, které relevantní části chybné struktury opraví na zcela správné (závislost i funkce).

Jako **neutrální zásah** označujeme změny, které chybnou strukturu změní, ale neopraví (nebo ne úplně) a přitom nezpůsobí závažné zhoršení struktury (např. změnu závislosti ze správné na chybnou). Mnoho **neutrálních zásahů** přinejmenším odstraní závažné rozpory s obecnými syntaktickými pravidly (např. odstraní situaci, kdy jsou na jednom slovese *být* závislé dva nekoordinované podměty, tak, že jeden označí za jmennou část verbonominálního predikátu, ovšem změní syntaktickou funkci u nesprávného podmětu, takže ve větě je sice už jen jeden podmět, ale ne ten správný).

Za **negativní, chybný zásah** považujeme jednak zhoršení chybné struktury (např. změna závislosti ze správné na chybnou), jednak jakoukoli změnu původně správně interpretované struktury. Pravidla byla koncipována tak, aby se chybným zásahům do správných struktur vyhnula i za cenu menšího počtu zásahů, ale struktury, s nimiž opravný program pracuje, jsou natolik variabilní, že nebylo vždy možno se závažné chybě vyhnout. Z principu je možné chybná pravidla vždy opravit a chyb se vyvarovat, uvedená úspěšnost odráží současný stav opravného programu.

Přehled zásahů vždy představujeme v tabulce, jejíž příklad zde uvádíme. Na prvním řádku je název pravidla, na druhém a třetím řádku jsou počty zásahů v korpusu SYN2005 přepočtené na 1 000 000 tokenů, na řádku čtvrtém a pátém jsou uvedena procenta úspěšnosti oprav (celkem, bez ohledu na typ změny).

Pravidlo pro opravu závislostí a funkcí u sloves se spojkou *-li*						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	295	226	0	0	0	521
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	88 %		12 %		0 %	

2.0.3 Rozdělení pravidel podle zaměření

Pravidla představíme ve čtyřech skupinách podle zaměření. První skupinu tvoří pravidla, jejichž cílem je opravit základní závislostní strukturu složené věty (jedna hlavní věta, věty vedlejší) či souvětí (více hlavních vět, případně věty vedlejší). Ve druhé skupině jsou obecná pravidla, která opravují chybné struktury uvnitř klauzí, ale nejsou přímo zaměřena na substantiva a jejich funkce. Třetí skupinu tvoří pravidla, jež opravují konkrétní chybné syntaktické funkce nebo závislosti

především syntaktických substantiv. Čtvrtá, nejpočetnější skupina se skládá z pravidel pro předložkové fráze, kde se použitý parser nejčastěji dopouští chyb ovlivňujících funkce či závislosti substantiv.

2.1 Pravidla korigující chybně určenou základní závislostní strukturu složených vět či souvětí

České psané texty se z velké části skládají ze souvětí a složených vět (v korpus SYN2005 připadá na větu v průměru cca 13 slov a 3 interpunkční znaménka). Struktury jsou často složité, obsahují větší množství interpunkce a spojek s různými funkcemi. Často se obtížně rozlišuje mezi větnou a členskou koordinací. Ve větách jsou navíc často elidované větné členy, slovesa, syntaktická substantiva aj., což zpracování vět dále komplikuje. V trénovacích datech je řídícím členem věty či souvětí jednou sloveso, jindy souřadicí spojka, interpunkce, nebo dokonce spojka podřadicí (samostatně stojící vedlejší věty). Pro parser je pak velmi obtížné správně zvolit řídící člen závislostní struktury a správně určit závislosti všech vět (hlavních i vedlejších), jež jsou reprezentovány slovesy.

Pět implementovaných pravidel se pokouší řešit některé typické chyby parseru. Kromě posledního pravidla, které je úzce zaměřené na jednu konkrétní strukturu, jsou pravidla poměrně obecná, odpovídají velkému množství různých potenciálně chybných struktur. V implementované verzi jsou pravidla velmi opatrná: nenajdou-li spolehlivé řešení chyby, raději chybnou strukturu ponechají nezměněnou. Pravidla jsou možná příliš opatrná, a jsou proto poměrně málo účinná.

2.1.1 Pravidlo pro opravu koordinace nekompatibilních členů, která se považuje za řídící člen závislostní struktury

Ve složitých větných strukturách obsahujících koordinaci a další interpunkci je pro parser někdy obtížné rozlišit, zda je koordinace větná nebo členská a zda slovesa ve větě patří do hlavních, nebo vedlejších vět. Parser v použitém nastavení neověřuje sousední uzly při přiřazování závislosti, dochází tedy k tomu, že na jedné koordinační spojce závisí ve výsledku syntaktická substantiva i slovesa.

Když se taková smíšená koordinace považuje za řídící člen závislostní struktury jako v příkladech (1), (2) a (3), mělo by ji toto pravidlo opravit (znak **#** označuje formální kořen věty, čísla vzdálenost řídícího uzlu v počtu slov, záporné číslo doleva).

(1) *Přikryl/**4** jsem/**-1** ji prostěradlem/**1** a/**Coord**/**#** bavlněnou dekou/**-2** .*

(2) *Sbírali/**13** rákosí a listy orobince , vrbové pruty , dlouhé tenké kořeny/**2** smrků a/**#** všechno/**-1** , co mohla/**-3** Ajla použít na pletení košů a provazů .*

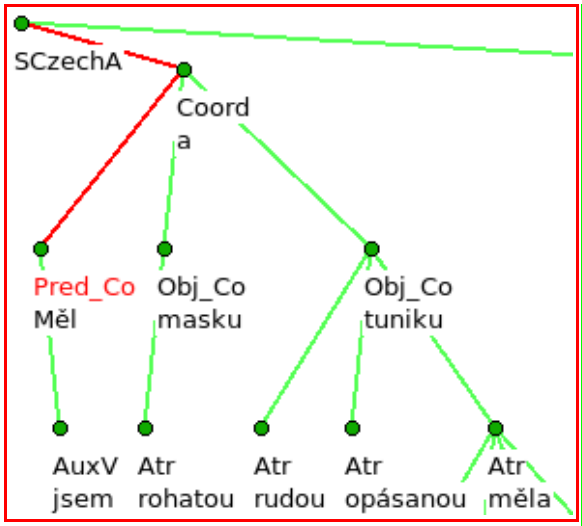
(3) *Nezval byl/**3** mimo prostor/**1** a/**Coord**/**#** čas/**-1** .*

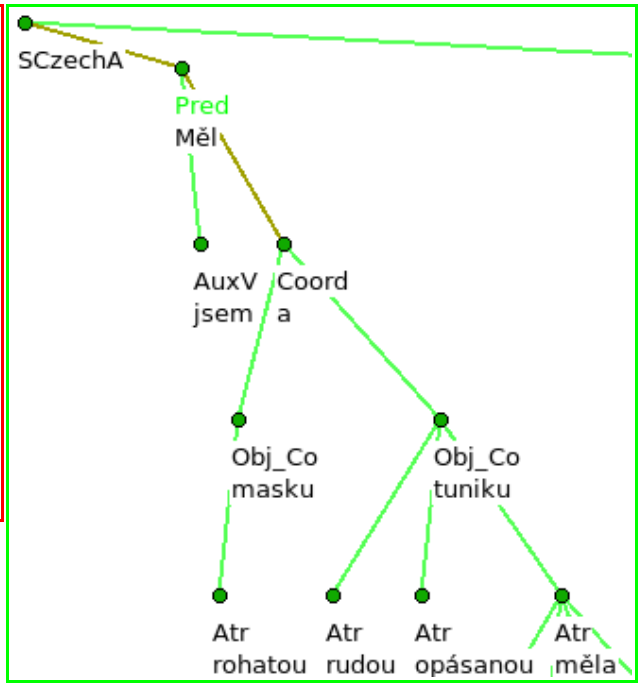
Pravidlo se aktivuje ke zpracování koordinace závislé přímo na formálním kořenu věty, na nichž jsou zároveň závislá alespoň dvě syntaktická substantiva a alespoň jedno sloveso v určitém tvaru. Pravidlo je opatrné, dokáže opravit jen jeden snadno řešitelný typ struktury: všechna substantiva závislá na koordinaci se shodují v pádě a na koordinaci je kromě toho závislé jedno sloveso neoznačené jako sloveso, které patří do vedlejší věty.

Za těchto podmínek opravné pravidlo označí pravděpodobné sloveso z hlavní věty (tj. sloveso, které je závislé na koordinaci a zřejmě nepatří do vedlejší věty) za řídící člen závislostní struktury (bude záviset přímo na formálním kořenu věty), pro koordinaci pak najde vhodný řídící uzel podle kontextu a pádu substantiv závislých na koordinaci: vhodným uzlem může být těsně předcházející předložka, blízké sloveso či předcházející substantivum, popř. sloveso, jež bylo dosud závislé na koordinaci a nyní je řídícím uzlem celé struktury.

(3b) *Nezval byl/**3**/**#** mimo/**-1** prostor/**Atr**/**Adv** a/**#**/**-2** čas/**Obj**/**Adv** .*

(4) *Měl/**4**/**#** jsem rohatou masku a/**#**/**-4** rudou opásanou tuniku , která měla vzadu přišitý kus provazu jako ocas .*





Pravidlo pro opravu koordinace nekompatibilních členů považované za řídící člen závislostní struktury						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	247	38	70	0	0	355
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	86 %		6 %		8 %	

2.1.2 Pravidlo pro slovesa ve vedlejších větách chybně považované za řídící člen závislostní struktury

Správné určení hlavní věty (a slovesa, které ji reprezentuje) je pro parser často velmi obtížné, znamená to zvolit z několika sloves jedno, které nepatří do vedlejší věty, přičemž větné struktury jsou někdy velmi složité, včetně vložených vedlejších vět aj. Vedlejší věta se ve formátu PDT, tedy v trénovacích datech, může projevit různým způsobem (sloveso závisí na podřadicí spojce nebo na slovesu závisí vztažné zájmeno či adverbium). Parser tedy v tomto určení často chybuje a označuje vedlejší věty za věty hlavní (1) nebo koordinuje vedlejší věty s hlavními (2).

(1) *Když/**AuxC**/**7** usoudila/**Adv**/**-1** , že/AuxC/-2 jsem se dost vynadíval/**Pred**/**#** , opět si blůzku zapnula/**Pred**/**-4** .*

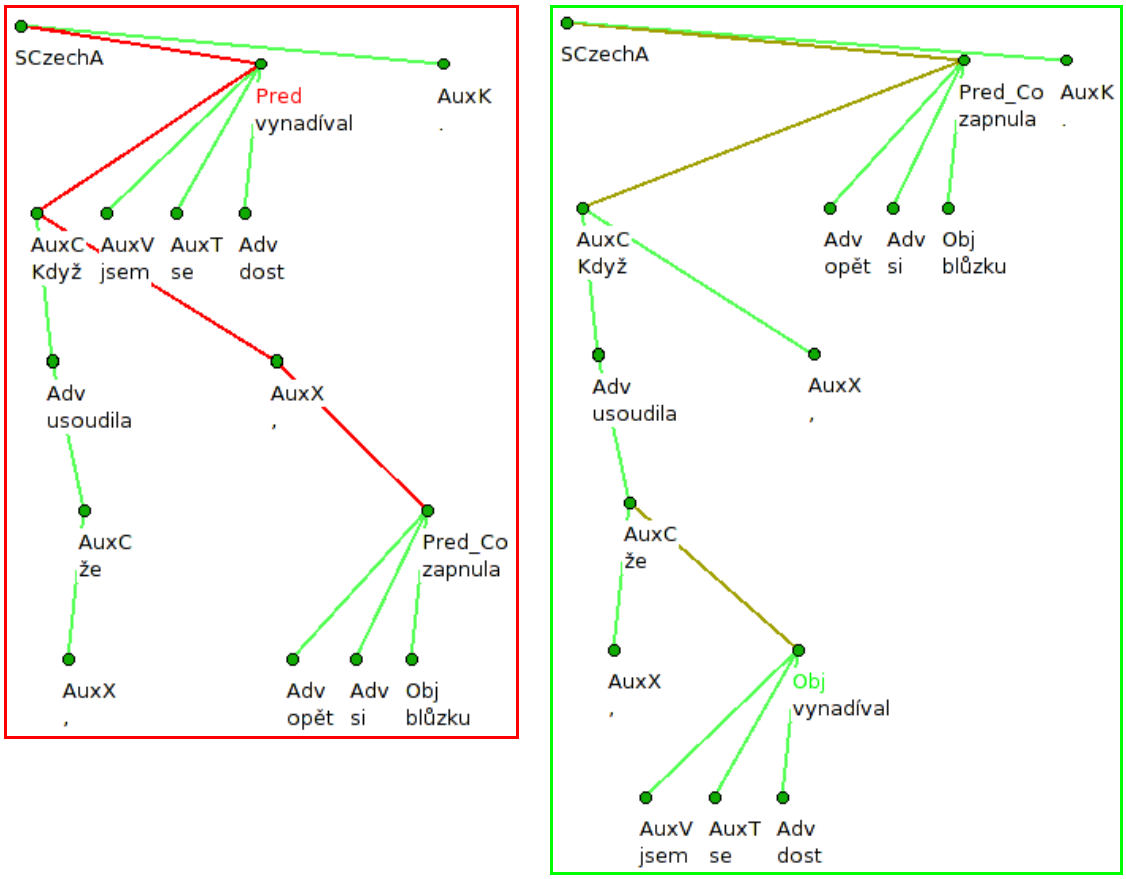
(2) *Snažil/**Pred**/**10** jsem se ti naznačit , abys/**4** ho zadržela/**Pred**/**2** , ale/**Coord**/**#** nechtěl/**Pred**/**-1** jsem křičet a vyplašit je .*

V základní části opravného programu je každému slovesu v určitém tvaru (nebo reprezentantu složeného slovesného tvaru, jímž může být např. infinitiv u složeného futura) přiřazen parametr, který rozlišuje, zda sloveso patří do hlavní nebo vedlejší věty (vztažné nebo s podřadicí spojkou). Když se pak v průběhu ověřování věty setká opravný program se slovesem, které podle přiřazeného parametru patří do vedlejší věty, ale parser ho interpretuje jako řídící člen celé závislostní struktury (sloveso závisí na formálním kořenu) nebo jako člen koordinace, která je řídícím členem závislostní struktury, spustí se toto opravné pravidlo.

Pravidlo rozděluje opravu na dvě větve: v jedné zpracovává struktury s vedlejší větou závislou přímo na formálním kořeni věty, ve druhé větvi zachází s koordinací označenou za řídící člen závislostní struktury, která obsahuje vedlejší větu. V první větvi pravidlo zjistí, zda se ve větě nachází jiné sloveso, které není označeno jako sloveso ve vedlejší větě, a je tedy pravděpodobně správným řídícím členem věty. Najde-li takové sloveso, změní ho na řídící člen věty. Pro vedlejší větu potom hledá vhodný řídící uzel: nejčastěji jím bude předcházející sloveso v určitém tvaru, tj.

slovo, které reprezentuje předcházející větu (1b).

(1b) *Když usoudila , že jsem se dost vynadíval , opět si blůzku zapnula .*



Ve druhé větvi pravidlo ověří parametry všech sloves, která jsou závislá na koordinaci. Pokud alespoň jedno z nich (lépe dvě) program nemá označené jako slovesa reprezentující vedlejší věty, stačí vyřadit vedlejší větu z koordinace a nalézt pro podřadicí spojku ve větě vhodný řídící uzel, přičemž sloveso má být závislé na podřadicí spojce. Vhodným řídícím uzlem je obvykle předcházející sloveso v určitém tvaru, i když v některých případech musí být pravidlo ve volbě řídícího uzlu sofistikovanější, bere v potaz například infinitivy závislé na modálních, fázových a dalších podobných slovesech (2b) a odkazovací obraty typu *na to, aby* (3).

(2b) *Snažil jsem se ti naznačit , abys/**4**/-2 ho zadržela/**Pred**/**Obj**/**2**/-2 , ale nechtěl jsem křičet a vyplašit je .*

(3) *Jsem/**13** už stará na to , abych/-**4**/-2 sviště**l**a/**6**/-1 po sjezdovkách jako kdysi , a/**#** nehodlám/-**1** se otravovat s nějakou novou technikou .*

Pravidlo pro slovesa ve vedlejších větách chybně považované za řídící člen závislostní struktury						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	134	96	29	0	0	259
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	76 %		15 %		9 %	

2.1.3 Pravidlo pro opravu závislostí souřadných hlavních vět

Musí-li parser pro řídící uzel celého souvětí volit mezi slovesem a souřadicí spojkou, často se rozhoduje chybně, někdy vytváří syntakticky nesmyslné kompromisní struktury. Příkladem takové struktury může být věta, kde je jedno sloveso závislé přímo na formálním kořeni, na něm je závislá koordinační spojka, na této spojce další sloveso. Ve skutečnosti měla být obě tato slovesa, popř. více sloves, koordinována, tj. řídícím uzlem struktury měla být spojka (1). Do koordinace jsou často chybně zapojeny další větné členy (2) a (3).

(1) *Byla/# hrozně pomalá a/-3 se vším se děsně babrala/-5 ;*

(2) *Anna obešla/# stůl/1 a/-2 chvíli/1 trvalo/-2 , než si ti dva všimli ,*

(3) *Každý Čech rozumí/# Slovákovi/1 a/-2 Slovák/1 rozumí/-2 Čechovi .*

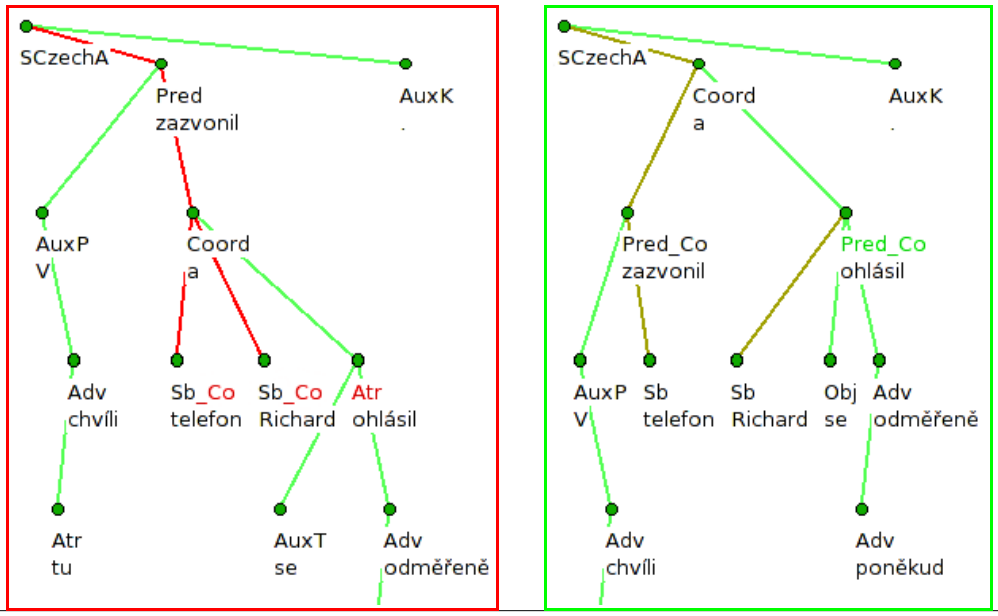
Pravidlo se spustí, když hlavní program při ověřování struktury věty narazí na sloveso označené jako řídící uzel celé závislostní struktury (visí na formálním kořeni věty), na tomto slovesu je závislá koordinační spojka, na této spojce je závislé sloveso, které nepatří do vedlejší věty (nejsou na něm závislá vztažná zájmena, v jeho klauzi nestojí podřadicí spojka).

Ani určení hranic klauzí (čárky, spojky aj., které oddělují klauze) na základě lingvistických pravidel není vždy zcela spolehlivé, protože se musí opírat o disambiguaci (s několika procenty chyb) a částečně i o závislostní strukturu vytvořenou parserem, kterou se snaží opravit. Aby nevnášelo do struktury další chyby, je tedy pravidlo opatrné a počítá pouze s jediným možným separátorem klauzí mezi dvěma hlavními slovesy: koordinační spojkou.

Pravidlo ověří, že mezi slovesem označeným jako řídící člen struktury a slovesem závislým na koordinační spojce nestojí kromě této spojky žádná jiná potenciální hranice klauzí. V takové konfiguraci pak označí spojku za řídící uzel celé struktury (#), původní hlavní sloveso bude závislé na této spojce. Mezi případnými dalšími členy koordinace pravidlo ponechá beze změny pouze další slovesa, která reprezentují hlavní věty. Slovesa, která reprezentují věty vedlejší (vztažné nebo uvozené podřadicí spojkou), budou převěšena na jiné vhodné uzly (vztažné na předcházející

antecedent, spojkové na předcházející sloveso, popř. na jedno z hlavních sloves ve větě). Ostatní větné členy (vyjma grafické znaky) pravidlo převěsí na sloveso, které je jim nejblíž: ty, které stojí vlevo od koordinace, na sloveso vlevo, ty, které stojí vpravo od koordinace, na sloveso na pravé straně. Pravidlo nemění syntaktické funkce jednotlivých větných členů, to ponechává na následujících pravidlech (4).

(4) *V tu chvíli zazvonil telefon a Richard se ohlásil poněkud odměřeně .*



Pravidlo pro opravu závislostí souřadných hlavních vět						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	1105	15	58	0	0	1078
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	96 %		4 %		0 %	

2.1.4 Pravidlo pro ověření závislostí a funkcí vztažných vět

Problémy s určením správné závislosti a funkce vztažných vět mají v zásadě stejné příčiny jako ostatní chyby týkající se celé větné struktury: složité struktury obsahující interpunkční znaménka, koordinace členské i větné aj. K tomu se přidávají obtíže s hledáním náležitého antecedentu vztažného zájmena dané mj. i tím, že mezi relativem a antecedentem stojí v závislostní struktuře sloveso (relativum je závislé na slovesu; sloveso, jež reprezentuje vztažnou větu, je závislé na antecedentu). Chyba může být méně závažného charakteru, kdy sloveso ve vztažné větě závisí na větném členu z předcházející věty, i když ne na správném antecedentu (1). Chyba v závislosti vztažných vět může také odrážet závažnější chybu v konstrukci závislostní struktury: vedlejší

vztažná věta může být například chybně považována za hlavní větu (2).

(1) *Ihned zpozoroval , že kromě klíče od vozu tam bylo několik dalších , které/**1** vypadaly/**Pred**/-**5** jako klíče od domu .*

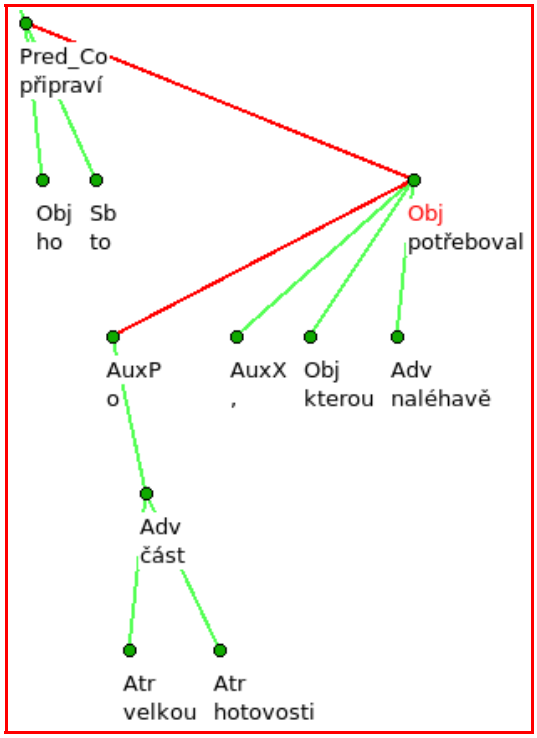
(2) *Tolik Shawovo oznámení/**5** o svatbě , které/**1** zaslal/**Pred**/**#** k otištění do londýnských novin 2 . června 1898 .*

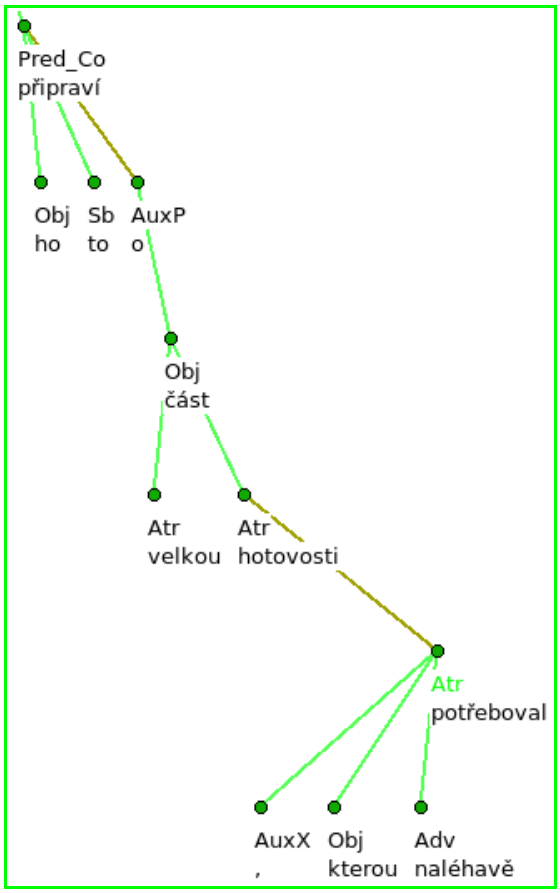
Pravidlo se zavolá pro každé sloveso, na němž je (přímo či nepřímo) závislé vztažné zájmeno. Před relativem musí stát čárka (popř. čárka a předložka), na vhodném místě před čárkou musí stát větný člen, který může sloužit jako antecedent vztažného zájmena (substantivum, některá zájmena, adjektiva, číslovky).

Pravidlo nejprve ověřuje možnou shodu relativa s antecedentem v rodě a čísle. Protože však disambiguace není zcela spolehlivá, nevyžaduje se shoda morfologických značek. Program ověřuje, že tvary relativa a antecedentu umožňují shodu v rodě a čísle, a to bez ohledu na přiřazené morfologické značky. Jestliže se relativum s antecedentem shoduje a sloveso není ve vztažné větě chybně označeno jako řídící člen závislostní struktury, opraví se závislost slovesa a případně i jeho syntaktická funkce.

Je-li však sloveso chybně určeno jako řídící uzel (3), je nejprve nutno vyhledat nový, správný řídící uzel pro celou závislostní strukturu a opravit podle toho závislosti ve větě. Tímto novým řídícím uzlem by mělo být v nejlepším případě sloveso (takové, které nestojí ve vedlejší větě, tj. v klauzi s daným slovesem není ani podřadicí spojka, ani vztažné zájmeno). Nenajde-li se takové sloveso, může jím být i antecedent relativa nebo jemu nadřazený větný člen.

(3) *Ale připraví ho to o velkou část hotovosti, kterou naléhavě potřeboval.*





Pravidlo pro ověření závislostí a funkcí vztažných vět						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	207	346	6	0	0	559
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	83 %		7 %		10 %	

2.1.5 Pravidlo pro opravu závislostí a funkcí u sloves se spojkou *-li*

Sloveso následované podřadicí spojkou *-li* je specifická struktura, která vyžaduje samostatné pravidlo. V PDT je (stejně jako u ostatních podřadicích spojek) spojka *li* řídícím uzlem celé věty, sloveso i spojovník závisí na spojce, spojka na slovese nadřazené věty. Struktura je poměrně častá, objevuje se cca třistakrát v trénovacích datech, přesto ji parser přibližně v polovině případů interpretuje chybně (závislosti nebo syntaktické funkce, často obojí). Někdy je jen prohozena závislost mezi slovesem a spojkou *li*, případně jsou chybně označené syntaktické funkce (1). Často je chybná širší závislostní struktura (2) a opravený příklad (3).

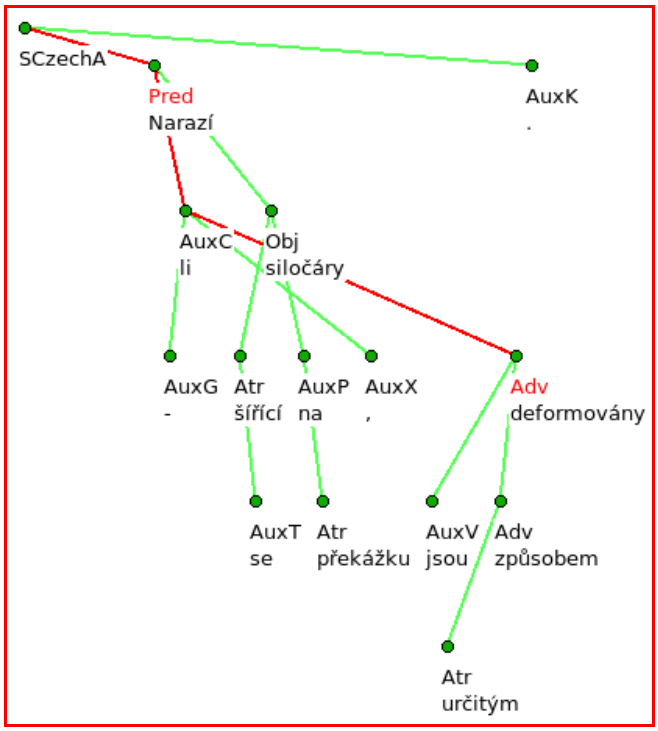
(1) *Je/**Obj**/**13** -/**1** li/-2 v čele odboru skutečně člověk pouze se základním vzděláním , porušila radnice zákon !*

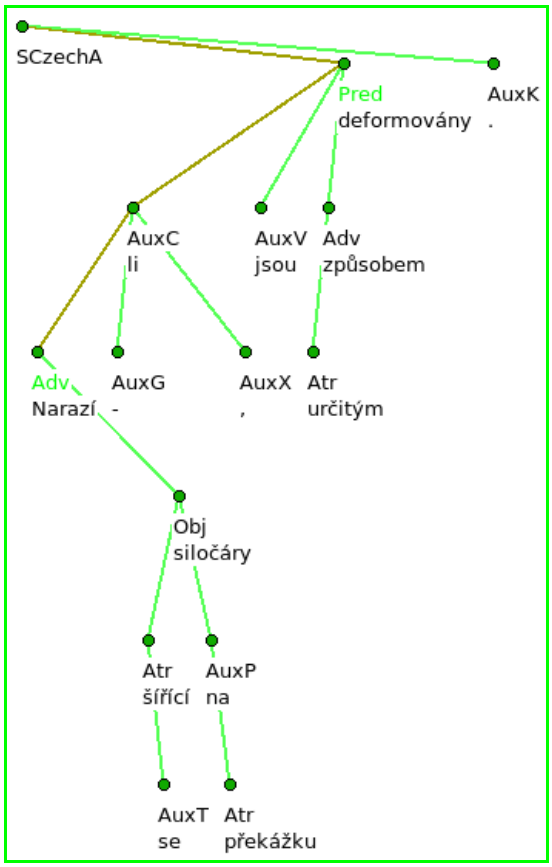
(2) *Vyšplhá/**Pred**/**2** -/**1** li/**# na konec druhého dílu , získává/**Adv**/**-6** 2 body .* (Oprava tohoto příkladu byla textově i graficky znázorněna v odstavci 2.0.2).

Pravidlo se volá vždy, když po slovese následuje spojovník a slovo *li*. Zároveň musí platit jedna z následujících podmínek: sloveso nebo spojovník nejsou závislé na spojce *li*, sloveso má chybnou syntaktickou funkci nebo je spojka *li* závislá přímo na kořeni věty a ve větě (souvětí) je více sloves.

Není-li sloveso závislé na spojce, změní pravidlo závislost slovesa na spojku s výjimkou pomocných sloves, která mají být závislá na hlavní části složeného slovesného tvaru (v tom případě se také ověří závislost druhé části složeného slovesného tvaru). Přitom se může také opravit závislost spojovníku, ale většinou to není třeba. Byla-li spojka závislá na předcházejícím slovese nebo přímo na kořeni věty, závislost se změní, pokud se ve větě najde jiné vhodnější sloveso (sloveso, v jehož klauzi nejsou podřadicí spojky ani vztažná zájmena).

(3) *Narazí-li šířící se siločáry na překážku, jsou určitým způsobem deformovány.*





Pravidlo pro opravu závislostí a funkcí u sloves se spojkou *-li*						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	295	226	0	0	0	521
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	88 %		12 %		0 %	

2.2 Obecná pravidla ověřující závislosti a syntaktické funkce větných členů

Skupina pěti obecných pravidel, která představíme v následujícím oddíle, je nesourodá. Spojuje ji jen to, že není primárně zaměřená ani na základní závislostní strukturu (složenou ze sloves a spojek) jako skupina předchozí, ani přímo na syntaktická substantiva nebo předložky (jako dvě následující skupiny).

První pravidlo opravuje závislosti větných členů, které chybně překračují hranici klauzí. Druhé pravidlo ověřuje závislosti a syntaktické funkce spojení se slovy *jako*. Třetí pravidlo se soustředí na zvratné zájmeno *se* a jeho syntaktické funkce. Poslední pravidlo, které řadíme do této skupiny, není vlastně samostatným pravidlem, ale pouze ověřuje, zda změny morfologických značek prováděných jinými pravidly nemají ovlivnit shodné přívlastky slov, u nichž změny probíhají.

2.2.1 Pravidlo pro větné členy, jejichž závislost chybně překračuje hranici klauzí

Jak bylo řečeno výše, MST parser nedokáže dobře pracovat s hranicemi klauzí. Někdy tak vytváří chybné struktury, v nichž je slovo závislé na řídícím větném členu za hranicí klauzí, přestože se přímo v klauzi nachází jiný, vhodný řídící uzel, neoddělený žádnou hranicí klauzí. Pravidlo se zaměřuje především na syntaktická substantiva, syntaktická adjektiva a předložky, které jsou závislé na slovese v sousední klauzi místo na slovese ve své klauzi.

V mnoha případech je zaváhání parseru pochopitelné, protože zkoumané slovo je od „svého“ slovesa odděleno jiným slovem, které se někdy vyskytuje na hranici klauzí: v příkladu (1) je to slovo *až*; popř. mezi slovesem a větným členem stojí neznámé slovo: v příkladu (2) slovo *Knightonem*. Často je ale taková chyba spíše překvapivá, protože hranice klauzí jasná je a správné řídící sloveso je blíže než to, na němž je zkoumané slovo chybně závislé: příklad (3) a opravené příklady (4) a (6).

(1) *tvrzení , že to ještě nedokazuje Baileyho verzi , podle které on narazil/-5 do oběti až jako/3 druhý , byla neudržitelná .*

(2) *Když byl na chvilku s Knightonem/X@--- sám/2 , poznamenal :*

(3) *Prosím tě/2 , nemluv v samých hádankách !*

Opravné pravidlo zaměřené na tento typ chybných konstrukcí patří k obecným, nespecializovaným opravným pravidlům. Vyhledává a opravuje struktury, v nichž je větný člen závislý na slovese, od kterého je oddělen nejméně jednou hranicí klauzí, přestože se v jeho klauzi nachází jiné sloveso v určitém tvaru, neoddělené od dotyčného žádnou hranicí klauzí. Změní závislost větného členu ze slovesa v jiné klauzi na závislost na nejbližším slovese v téže klauzi, případně také změní syntaktickou funkci.

Pravidlo se spouští na předložky, syntaktická substantiva a adjektiva, jejichž závislost překračuje hranici klauzí. Nalezne-li pravidlo jiné sloveso v určitém tvaru, které od zkoumaného slova není odděleno hranicí klauzí, je závislost zkoumaného slova určena chybně, pravidlo se pokusí vyhledat pro zkoumané slovo vhodný řídící větný člen (může jím být nalezené sloveso nebo jiný vhodný větný člen z klauze podle kontextu, slovního druhu a pádu slova, jehož závislost pravidlo mění, popř. podle valence okolních slov).

(4) *Tento postřeh si zapamatujme , vrátíme se k/4/-2 němu totiž a začleníme jej do širšího kontextu později*

(5) *Tři motory se zážehovým paprskem (celkem mají být čtyři/2/-2) dávají elektrický výkon 320 kW .*

(6) *Pórky očistíme a světlou část/-3/1 nakrájíme na dílky .*

Pravidlo pro větné členy, jejichž závislost chybně překračuje hranici klauzí						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	550	15	0	0	0	565
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	91 %		7 %		2 %	

2.2.2 Pravidlo pro spojení se spojkami *jako*

V PDT jsou obraty se slovy *jako* značkovány různě podle významu. Příklady pocházejí z manuálu PDT[2]. Ve srovnávacích výrazech má *jako* funkci spojky ***AuxC***, srovnávací výraz po spojce *jako* je závislý na spojce a má obvykle funkci ***ExD*** (většinou lze výraz interpretovat jako elipsu predikátu): *spí jako/**AuxC**/**-1** zabitý/**Exd**/**-1***. Jinak má slovo *jako* syntaktickou funkci ***AuxY***

2 http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/html/ch03s07s09.html

(což je kategorie zahrnující několik různých funkcí jako součást víceslovné spojky či částice, „pokleslá vsuvka“ aj.) a je závislé na následujícím substantivu či adjektivu (předmětu či doplňku): *odmítl nabídku jako/**AuxY/1** málo atraktivní/**Atv/-3** ; připadá mi to jako/**AuxY/1** neskutečné/**Obj/-4***.

Toto pojetí je z lingvistického hlediska možná opodstatněné, ale pro stochastický parser je nevhodné, protože často závisí na významových nuancích, parserem nerozeznatelných. Rozlišení mezi doplňkem a předmětem možné je, předměty může parser rozeznávat podle sloves, např. *označit*, *chápat*, *vnímat*, i když v trénovacích datech je takových předmětů málo (cca 30). V důsledku složité situace chybuje parser u spojek *jako* a *než* a u slov po nich následujících velmi často. Někdy jen chybně určí funkci a závislost substantiva (adjektiva) po spojce (1), (4) a (5), jindy chybně interpretuje celou strukturu: jak spojku, tak následující slovo (2) a (3).

(1) *Snažil se vypadat jako/**AuxY/1** mírotvůrce/**Atv/-3** , zatímco v podstatě vytvářel příznivé podmínky*

(2) *možná i proto si ho jako/**AuxY/2** horký brambor/**Obj/1** přehazuje z jednoho okrsku na druhý*

.

(3) *boj lidí proti strojům , z něhož nevyhnutelně vycházeli lidé jako/**AuxC/-2** poražení/**ExD/-1** .*

Pravidlo se nesnaží rozlišit automaticky mezi srovnávacími obraty na jedné straně a doplňky a předměty na druhé, tj. nerozhoduje mezi ***AuxC*** a ***AuxY***, jen napravuje závažné chyby v rámci těchto kategorií. Pravidlo se zavolá pro všechny případy nesouladu mezi funkcí a závislostí spojky a substantiva / adjektiva, které k ní patří: spojka je označena jako ***AuxY*** a je závislá na následujícím slově, které nemá funkci ***Obj***, ***Atv*** nebo ***AtvV***; spojka je označena jako ***AuxC*** a je na ní závislé následující slovo, které nemá funkci ***ExD*** nebo ***Adv***.

Pravidlo uvádí do souladu syntaktickou funkci substantiva či adjektiva s funkcí spojky, kterou nemění. Je-li spojka označena ***AuxY***, zvolí mezi ***Obj*** a ***Atv/AtvV*** podle řídícího uzlu, funkci ***Obj*** může dostat pouze slovo závislé na slovesech či adjektivech z odpovídajícího seznamu (*chápat*, *jevit se*, *označit*; *označovaný*, *sloužící*), ostatním bude přiřazena funkce ***Atv*** (pokud je slovo závislé na substantivu) nebo ***AtvV*** (pro slova závislá na slovesech). Je-li spojka označena jako ***AuxC***, přiřadí se substantivu, které je na ní závislé, funkce ***ExD***, adjektivu funkce ***Adv***. Spolehlivý algoritmus pro automatickou volbu mezi označením spojky ***AuxY*** a ***AuxC*** nebyl bohužel nalezen, pravidlo tak často pouze mění syntaktickou funkci substantiva ze zcela nepřípustné na jinou, ale bohužel také nesprávnou.

(4) *čímž chtěl naznačit , že jako/**AuxY** obyčejný voják/**Sb**/**AtvV**/**1** není odpovědný ,*

(5) *Miluj svého bližního jako/**AuxC**/**-1** sám/**AtvV**/**ExD**/**-1** sebe .*

Pravidlo pro spojení se spojkami *jako*						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	0	1	403	0	0	404
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	51 %		49 %		0 %	

2.2.3 Pravidlo pro určení syntaktické funkce reflexiva *se*

Reflexivu *se* jsou ve formalismu PDT přiřazeny čtyři různé syntaktické funkce podle jejich použití ve větě: ***Atr*** pro *se* závislé na deverbativním substantivu; ***AuxT*** pro *se* závislé na slovese reflexivu tantum (*smát se*), popř. na slovese, jež má znatelně odlišný význam jako zvratné (*hodit se* vs. *hodit)*; ***AuxR*** pro reflexivní pasivum (*Tyto změny se/**AuxR** v roce 1915 uskutečnily postupně všude*) a ***Obj*** tam, kde plní zájmeno funkci předmětu (*Ministr Škromach se/**Obj** už teď chytá za hlavu*). Rozpoznání těchto funkcí je však pro parser velmi obtížné, při určování funkcí je často nutné opřít se o vlastnosti sloves (z trénovacích dat například nelze získat dostatečná data o reflexivech tantum či o netranzitivních slovesech), o strukturu věty či o význam slov. Syntaktické funkce slova *se* parser tedy často určuje chybně.

Dosud implementované pravidlo opravuje jen chybně přiřazenou funkci ***AuxT***, nepůsobí v opačném směru, nevyhledává aktivně reflexiva *se* pro zvratná slovesa ani neopravuje funkci slova *se* v sousedství deverbativních reflexivních substantiv, i když by to bylo vhodné (např. ve spojení *ztotožnění se* je tvar *se* správně interpretován jako přívlastek substantiva pouze asi v jedné třetině výskytů v korpusu). V budoucnu s doplněním těchto oprav do pravidla počítáme.

Pravidlo se spustí, když hlavní program narazí na slovo *se* označené ***AuxT***, které je závislé na slovese, jež není reflexivum tantum (ani sloveso typu *hodit se*, *domluvit se*...). Je-li slovo *se* závislé na slovese *být* nebo na modálním či fázovém slovese, pokusí se pravidlo najít v klauzi jiné zvratné sloveso, jemuž by mohlo slovo *se* přiřadit. Pokud se to nepodaří nebo slovo *se* nezávisí na takovém slovese, změní pravidlo jeho syntaktickou funkci na ***Obj*** nebo ***AuxR*** podle kontextu a vlastnosti slovesa: je-li sloveso netranzitivní, je v reflexivním pasivu a slovo *se* má mít funkci ***AuxR*** (1). Je-li sloveso tranzitivní a je v první či druhé osobě, má *se* funkci předmětu. Tranzitivní sloveso ve třetí osobě s neurčeným rodem nebo mužského životného rodu také pravděpodobně není reflexivní

pasivum, slovo *se* pak bude mít funkci ***Obj*** (2). U ostatních rodů je rozhodování obtížné, častěji je *se* součástí reflexivního pasiva než předmětem, podle toho pravidlo upravuje funkce (3) a (4). Podíl změn, které funkci v konečném důsledku neopraví, je poměrně vysoký (5): oranžově označená funkce byla opravena nevhodně.

(1) *Pěšími vojáky se/**AuxT**/**AuxR** jako pomocnými sbory pohrdalo .*

(2) *Všichni jsme se/**AuxT**/**Obj** namačkali do svatyně předků proti těm dvěma portrétům*

(3) *na neutrálních lodích budou považovány za kontraband , pokud se/**AuxT**/**AuxR** neprokáže opak .*

(4) *Vyjednávání se/**AuxT**/**AuxR** protahovala a díky publicitě , které se jim dostalo , ještě zvyšovala*

(5) *část britské královské rodiny se/**AuxT**/**AuxR** přejmenovala z Battenbergů na Mountbatteny .*

Pravidlo pro určení syntaktické funkce reflexiva *se*						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	60	12	4351	0	0	4423
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	86 %		12 %		2 %	

2.2.4 Pravidlo pro promítnutí změny morfologické značky substantiva na jeho shodné přívlastky

Uvádíme zde i dílčí pomocné pravidlo, které samo neopravuje žádnou chybnou strukturu, ale jen promítá změnu morfologické značky z řídícího substantiva na jeho shodné přívlastky. Protože však změny provedené tímto pravidlem evidujeme samostatně, hodnotíme samostatně i úspěšnost změn, které se provedou jeho prostřednictvím.

Pravidlo se zavolá vždy, když se jiné pravidlo chystá změnit morfologickou značku syntaktického substantiva. Ověří, zda tvary všech shodných přívlastků syntaktického substantiva jsou homonymní tak, aby se i jejich morfologická značka mohla změnit (pád, rod, číslo). Pokud homonymní nejsou, změna je zablokována a neprovede se ani v pravidle, které původně změnu iniciovalo. Pokud homonymní jsou, změna se provede a zároveň je umožněna i změna morfologické značky řídícího syntaktického substantiva. Úspěchy i chyby tohoto pravidla plně odpovídají úspěchům a chybám nadřazeného pravidla, které je využívá.

V příkladech (1), (2) a (3) nadřízené pravidlo opravuje struktury se dvěma nekoordinovanými subjekty závislými na jednom slovese. Příklady (1) a (2) ukazují správnou změnu, příklad (3) změnu chybnou (modře původně správná morfologická značka, oranžově chybná oprava).

(1) *Historici předkládající tato/**PDNP1**/**PDNP4** tvrzení připouštějí , že vzrůstající síla*

(2) *Včely čalounice vykusují polokruhovité/**AAIP1**/**AAIP4** kousíčky listů růží a vystýlají jimi buňky pro larvy .*

(3) *Velké kusy horniny drtí mamutí/**AAIP1**/**AAIP4** drtiče do velikosti tenisových míčků .*

Pravidlo pro promítnutí změny morfologické značky substantiva na jeho shodné přívlastky						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	0	0	0	0	121	121
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	68 %		0 %		32 %	

2.3 Pravidla pro opravu závislostí a syntaktických funkcí syntaktických substantiv

Cílem této práce je získat co nejspolehlivější data o syntaktických funkcích substantiv, většina implementovaných pravidel je proto zaměřena na jejich ověření a opravu. Parser nejčastěji chybuje v předložkových frázích, pravidla, která je opravují, budou představena až v následujícím oddíle. Mnoho chyb spočívá ale také v určení závislostí a syntaktických funkcí bezpředložkových pádů. Často je jejich příčinou i chybná disambiguace, kterou se pravidla s větším či menším úspěchem také pokoušejí opravit. Opakovaně se také objevují struktury, které se v trénovacích datech vůbec nemohly objevit a které vyplývají z omezení použitého parseru (nemožnost ověření sesterských uzlů aj.), například na jednom slovese často závisejí dva nekoordinované subjekty (***Sb***) nebo jmenné části verbonominálního přísudku (***Pnom***), dvě různé syntaktické funkce jsou koordinované (formálně závislé na jedné koordinaci) atd. Do této skupiny patří šest opravných pravidel, uvádíme je v pořadí podle počtu zásahů v korpusu.

2.3.1 Pravidlo pro opravu shodných substantivních přívlastků

V PDT se souřadné struktury typu *pan Novák* považují za spojení shodného substantivního přívlastku s řídícím jménem: *pan* je přívlastek, *Novák* je řídící substantivum. V případě složitější struktury (více apelativ, jméno a příjmení: *pan ministr zemědělství Jan Fencl*) jsou všechna shodná substantiva ve skupině závislá na posledním: *pan/**Atr**/**4** ministr/**Atr**/**3** zemědělství/**Atr**/**-1** Jan/**Atr**/**1** Fencl*. Je to arbitrární rozhodnutí, jak v jednoduché závislostní struktuře zacházet s těmito složitými

skupinami. Protože se však v trénovacích datech objevuje jen omezený počet takovýchto struktur, který neumožňuje jejich bezpečné rozpoznání, parser zde poměrně často chybuje, a to v obou směrech: neznačkuje substantiva ve vhodných strukturách jako shodné přívlastky (1) a značkuje substantiva jako shodné přívlastky ve strukturách, které takto interpretovat nelze (2).

(1) *na nástěnce vedle nejšpinavějšího portrétu předsedy/**Atr**/**-1** Maa/**Atr**/**-1** , jaký jsem kdy viděl .*

(2) *během těch tří dnů/**Atr**/**1** mise/**Adv**/**-4** postoupili Američané více než o dvacet kilometrů*

Pravidlo se spouští ve dvou případech, jimž také odpovídají větve řešení. V prvním případě je substantivum závislé na následujícím substantivu (ne nutně těsně následujícím), a to přímo, ne přes spojku či předložku, ale není to proprium mužského životného nebo ženského rodu ani apelativum ze seznamu cca 1200 substantiv, které se v takovýchto strukturách vyskytují (*pan*, *docentka*, *ředitel*, *sopranistka*...), popř. se neshoduje v rodě či pádě.

Ve druhém případě vedle sebe stojí dvě substantiva v singuláru, shodují se v rodě i pádě, první patří do výše zmíněné kategorie apelativ (typických „shodných substantivních přívlastků“) nebo je to křestní jméno (ze seznamu cca 900 jmen), druhé je také apelativum z téhož seznamu, křestní jméno nebo příjmení (ze seznamu cca 2000 jmen). První substantivum není závislé na druhém ani na dále následujícím substantivu stejného typu. Druhý případ je tedy ve svých požadavcích restriktivnější, požaduje splnění více podmínek, aby pravidlo do struktury zasáhlo a označilo sousedící substantiva za strukturu se shodným substantivním přívlastkem.

Pravidlo pak ve dvou větvích ověřuje a opravuje pravděpodobně chybné shodné substantivní přívlastky a struktury, které by pravděpodobně měly být označeny jako shodné substantivní přívlastky.

2.3.1.1 Pravděpodobně chybně určené shodné substantivní přívlastky

Vstupem této větve pravidla jsou pravděpodobně chybně určené substantivní přívlastky. Možné řešení záleží na pádu substantiv a jejich okolí, tedy zda pravidlo dokáže nalézt vhodné řídící uzly pro substantiva. Pro první substantivum (dosud závislé na následujícím) je nutné najít nový řídící uzel, může to být původní řídící uzel druhého substantiva, sloveso v klauzi s odpovídající valencí nebo chybějícím subjektem, předcházející substantivum či předložka aj.

Pokud je druhé substantivum v genitivu nebo pokud odpovídá jmenná skupina shodnému typu *město Praha* (kde se v PDT řeší závislost opačně) nebo je druhé substantivum v nominativu a první slovo umožňuje následující nominativ jmenovací (*řekou Indus*), označí pravidlo druhé substantivum

za přívlastek ***Atr*** prvního substantiva (3).

Když se první a druhé substantivum neshodují v pádě a druhé může být podmětem či předmětem slovesa v klauzi, změní se závislost druhého substantiva na blízké sloveso spolu se změnou funkce (4).

Není-li nic z toho možné (pro první ani pro druhé substantivum), žádná oprava se neprovede. Pravidlo se nepokouší o opravu případně chybné morfologické značky, na to je struktura příliš složitá a variabilní.

(3) , *blízko byla řeka/**Atr**/**Sb**/**1**/**-1** Nežárka/**Sb**/**Atr**/**-2**/**-1** s mlýnem a do města vedly staré , bohaté aleje .*

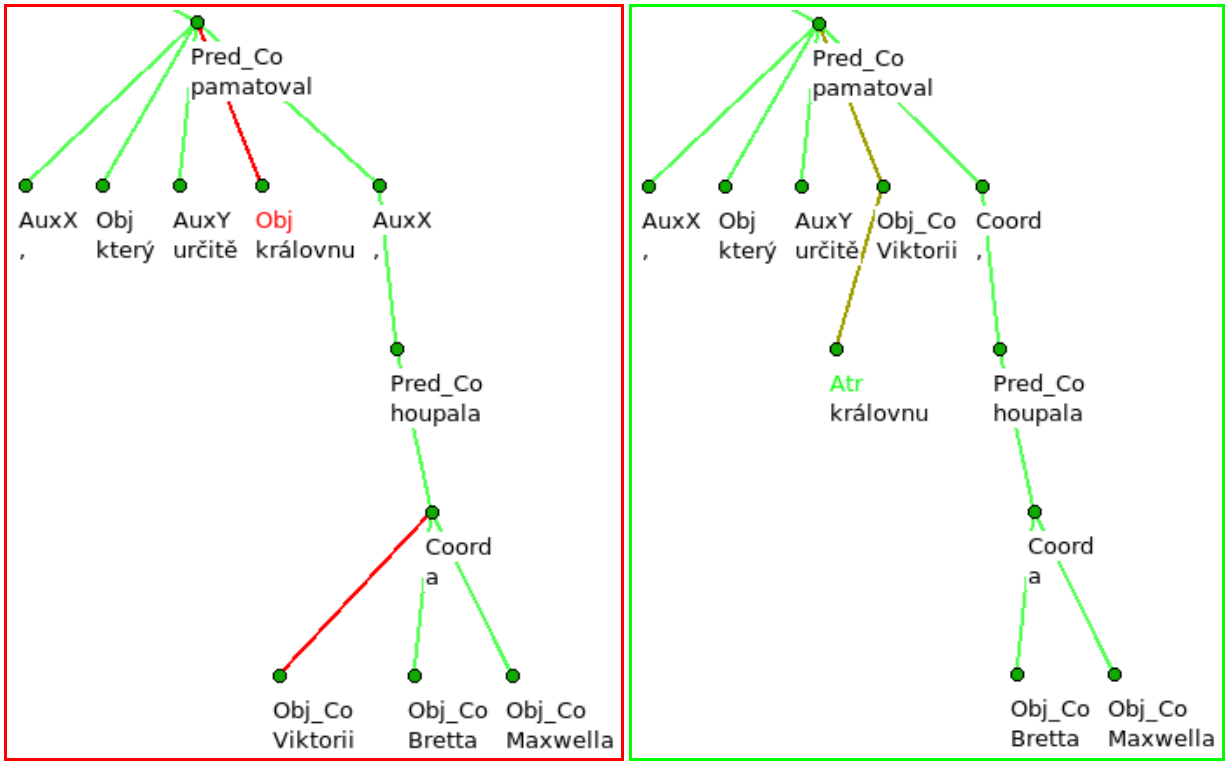
(4) *Když minuli vlnolam , navedl několika jistými posunky/**Atr**/**Obj**/**1**/**-3** člun/**Obj**/**-4** ke královskému molu*

2.3.1.2 Pravděpodobné shodné substantivní přívlastky neurčené jako takové

Vstupem druhé větve pravidla jsou vedle sebe stojící substantiva (nebo oddělené přísně vymezenými neshodnými přívlastky jako v následujících skupinách: *mistr světa Tomáš Dvořák*, *ministr zahraničí David Levy*), jež se shodují v pádu, rodu i čísle a patří do výše zmíněných skupin apelativ či proprií. První ze substantiv (a postupně všechna další) by mělo být závislé na posledním a mít syntaktickou funkci ***Atr***. Pro druhé (poslední) substantivum je třeba zvolit vhodný řídící uzel. Tím může být řídící uzel prvního (5) a (6) nebo druhého substantiva (pokud druhé nebylo závislé na prvním), jestliže splňují podmínky (stejná klauze, vhodná kombinace slovesa a substantiva z hlediska pádu, valence atd., pád a pozice předložky atd.). Není-li ani jeden z původních řídících uzlů vhodný, pokusí se pravidlo najít jiný řídící uzel v nejbližším okolí. Jestliže vhodný řídící uzel nenajde, opraví se jen závislost prvního substantiva na druhé a druhé převezme řídící uzel prvního (přestože tak konstrukce zůstane pravděpodobně chybná, opraví se jen její vnitřní struktura).

(5) *Zjevně zaskočen byl z včerejších myšlenkových obratů ministra/**Atr**/**-1**/**2** obrany/**Atr**/**-1** Tvrdíka/**Atr**/**-1**/**-3** premiér Vladimír Špidla .*

(6) *Postávala před školou a v kočárku , který určitě pamatoval královnu Viktorii , houpala Bretta a Maxwella .*



Pravidlo pro opravu shodných substantivních přívlastků						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	682	373	8	0	0	1063
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	81 %		16 %		3 %	

2.3.2 Pravidlo pro opravu dvou subjektů závislých na jednom slovese

Parser při přiřazování uzlů ve struktuře, kterou vytváří, nemůže ověřovat sousední uzly, pouze uzly nadřazené a podřazené. Častým typem chyby parseru je tak struktura, v níž jsou na jednom slovese závislé dva nekoordinované subjekty, parser totiž nemůže ověřit, zda již danému slovesu jeden subjekt nepřiřadil. Oprava této snadno identifikovatelné chyby je však relativně složitá, podrobnou analýzou věty je totiž třeba určit, které z mnoha možných řešení je nejlepší. Poměrně často je chyba způsobena chybným morfologickým značkováním: i tu může pravidlo odstranit. V současné verzi opravného programu je sedm možností, jak nalezenou chybu opravit (ne vždy úspěšně a ne každou takovou chybu je program v současnosti schopen opravit, přestože ji identifikuje).

Pravidlo se spustí, když hlavní program narazí na sloveso, na němž jsou závislé dva

nekoordinované subjekty. Pravidlo potom vyhodnocuje charakteristiku slovesa a pozici subjektů ve větě a podle toho volí větev, v níž se pokusí o nápravu chybné struktury. Postup a příklady řešení uvádíme u jednotlivých větví řešení.

2.3.2.1 Dva subjekty závislé na tranzitivním slovese, oba uvnitř klauze

Nejčastější podtyp chyby má příčinu v chybném morfologickém značkování, popř. v textové chybě (2): jedno substantivum má chybně určený pád, nominativ místo akuzativu. Parser na základě chybných dat určil chybně funkci (***Sb*** místo ***Obj***):

(1) *Neštěstí/**NNNS1**/**Sb** také prověřuje Inspektorát/**Sb** bezpečnosti práce ,*

(2) *Jakmile byla 28 . červenec/**NNNS1**/**Sb** vyhlášena válka/**Sb** mezi Rakouskem a Srbskem*

Tato větev opravy je zvolena, pokud jsou oba subjekty ve stejné klauzi jako sloveso a alespoň jeden ze subjektů je pádově homonymní (akuzativ – nominativ). Dále musí být sloveso v činném rodu a na slovese nesmí být závislý jiný předmět ve čtvrtém pádě ani reflexivum *se* (3), nebo musí být substantivum vyjádřením míry nebo času (4). Za těchto podmínek změní program morfologickou značku (z nominativu na akuzativ) i funkci homonymního substantiva (ze ***Sb*** na ***Obj***, popř. ***Adv***).

(3) *Důležitost/**NNFS1**/**NNFS4**/**Sb**/**Obj** těchto pokusů vědci/**Sb** zdůrazňují i tím , že dospívající mládež*

(4) *Celé následující století/**NNNS1**/**NNNS4**/**Sb**/**Adv** byl stroj/**Sb** v Marly považován za další div světa .*

Jsou-li obě substantiva homonymní a mohou se přitom shodovat se slovesem v rodě i čísle, opraví pravidlo značku a funkci **druhého** substantiva, kde je o něco vyšší pravděpodobnost úspěchu, nicméně cca ve 30 % případů tak strukturu neopraví správně, ale zhorší celkovou disambiguaci věty (5).

(5) *Význam/**NNIS1**/**Sb** má však i prostředí/**NNNS1**/**NNNS4**/**Sb**/**Obj** jeslí , školy a zaměstnání , prostředí v době volna .*

2.3.2.2 Dva subjekty závislé na slovese *být*, oba uvnitř klauze

Jsou-li na jednom slovese *být/bývat* závislé dva subjekty (v nominativu) a na slovese přitom není závislý jmenný přísudek, lze předpokládat, že jeden ze subjektů by správně měl být označkován jako jmenný přísudek. Je-li jedním ze subjektů zájmeno *to*, bude jako jmenný přísudek označen druhý subjekt bez ohledu na shodu se slovesem (6), jinak bude opravena funkce u toho syntaktického substantiva, které se neshoduje se slovesem v rodě a čísle (7). Shodují-li se obě syntaktická substantiva se slovesem, bude opravena funkce u druhého substantiva v pořadí, opět nejde o opravu jistě správnou, ale pravděpodobněji správnou (8).

(6) *Bude to/**Sb** asi něco/**Sb**/**Pnom** o kytičkách , myslela jsem si*

(7) *Byli ti tvorové vůbec ještě ženy/**Sb**/**Pnom** ?*

(8) *On/**Sb** je totiž ten typ/**Sb**/**Pnom** člověka , který může přemýšlet , jenom když mluví .*

2.3.2.3 Dva subjekty těsně vedle sebe, pojmenování osoby

Tvoří-li dva nekoordinované subjekty stojící ve větě těsně vedle sebe dvojici, jež se shoduje v rodě, čísle i pádě, a oba subjekty jsou substantiva, a to buď vlastní jména, nebo substantiva ze seznamu generických označení osob (jako *pan, paní, soudruh, doktorka, prezident* aj.), musí být v rámci formalismu PDT první z nich označen jako shodný přívlastek závislý na druhém jménu (podrobně se těmto skupinám věnuje pravidlo 2.3.1). Jsou-li obě substantiva v nominativu, parser někdy nedokáže správně identifikovat jejich syntaktickou funkci a závislost, obě zavěsí na sloveso a oběma přiřadí funkci ***Sb***. Správnou opravou je v tomto případě změna závislosti i funkce prvního substantiva (9) a (10).

(9) *Za Davidem Moravcem vyrazila jeho dívka/**Sb**/**Atr**/**-2**/**1** Jana/**Sb** .*

(10) (*slyšel ji říkat , jak moc se jí líbila Anna/**Sb**/**Atr**/**-1**/**1** Kareninová/**Sb***)

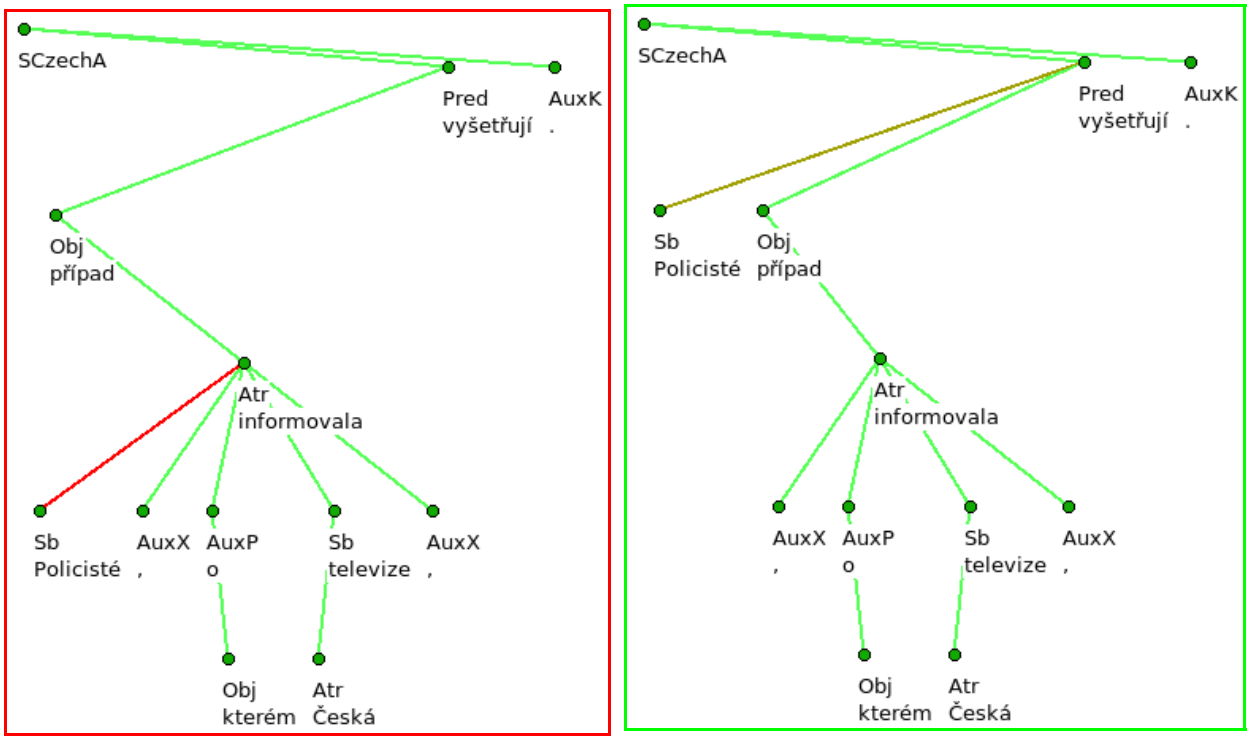
2.3.2.4 Dva subjekty závislé na jednom slovese, jeden přes jednu hranici klauze

Je-li větný člen závislý na slově, jež patří do sousední klauze, je závislostní struktura pravděpodobně chybná (ve formátu PDT kromě spojek a sloves, které v souvětích reprezentují celé vedlejší věty). Závislost může přecházet přes hranice klauzí jen tehdy, když mezi větnými členy stojí vnořená vedlejší věta, jež by sama měla být ukončena další hranicí klauze. Pokud tedy vazba jednoho ze subjektů v identifikované chybě přechází přes právě jednu hranici klauzí a tento subjekt

nemá ve „své“ části věty žádné sloveso, je třeba hledat jiné, vhodnější sloveso, na kterém by měl být subjekt závislý. Jinak je třeba hledat řídící sloveso po vložené vedlejší větě. Takové sloveso nesmí být samo ve vnořené vedlejší větě (musí tedy následovat po čárce, po níž nestojí hypotaktická spojka ani vztažné zájmeno či příslovce), nesmí mít jiný subjekt a musí se se subjektem, pro nějž hledáme řešení, shodovat v čísle i v rodě (11 a 12). Nalezne-li opravný program takové sloveso, může na něj subjekt převěsit, a chybu se dvěma subjekty tak opravit, jako v následujícím příkladu, kde sice nebyla opravena celá chybná struktura, ale chyba se dvěma subjekty byla odstraněna.

(11) *Přesná doba/**Sb**/4/8 , kdy posily/**Sb** dorazí , zatím však není známa , neboť si budou*

(12) *Policisté případ , o kterém informovala Česká televize , vyšetřují .*



2.3.2.5 Dva subjekty závislé na slovese *být*, jeden subjekt přes dvě hranice klauzí

Je-li na jednom slovese *být/bývat* závislý jeden subjekt uvnitř klauze a druhý přes dvě hranice klauzí, přičemž vzdálený subjekt není ve vedlejší větě (uvozené hypotaktickou spojkou či vztažným zájmenem nebo příslovcem), pravděpodobně jsou správně určeny závislosti, ale chybně určena funkce. Tato větev pravidla změní syntaktickou funkci druhého subjektu na ***Pnom***, bez ohledu na shodu. Shoda u dvou nominativů závislých na slovese *být* nerozhoduje, viz (13) a (14).

*(13) A že tým/**Sb** , který ženu operoval , nebyli žádní mladí cucáci/**Sb**/**Pnom** ,*

*(14) Jediná voda/**Sb** , kterou v bytě našli , byl nepatrný doušek/**Sb**/**Pnom** v odtokové trubce pod*

dřezem .

2.3.2.6 Osamostatněné přívlastky

Část nadbytečných subjektů ve strukturách se dvěma subjekty jsou chybně oddělené přívlastky (shodné i neshodné), jež parser zavěsil přímo na sloveso a přiřadil jim funkci ***Sb***. Parser takto zachází se slovy, která často stojí samostatně; takové rozhodnutí má tedy určitou oporu v trénovacích datech. Opravné pravidlo identifikuje tuto chybu u některých spojení s čísly, u vztažných zájmen a za přísně omezených podmínek také u ukazovacího zájmena *ten*.

Číslicí zapsaná čísla označená jako ***Sb***, která jsou závislá na slovese, na němž je závislý i jiný subjekt, a jež těsně následují po několika typických substantivech často rozvíjených číslovkou psanou číslicí (*rok*, *číslo*, *Praha*, *strana* aj., dále zkratky jako *odst.*, *čl.*, *str.*), označí tato větev pravidla jako ***Atr*** závislý na předcházejícím substantivu (15).

Vztažná zájmena přivlastňovací stojící před substantivem, s nímž se mohou shodovat v rodě, čísle i pádě, se mohou vždy převěsit na substantivum se změnou funkce (16). U zájmen *který* a *jaký* lze spolehlivě provést změnu, jen když před zájmenem stojí čárka a před ní sloveso, takže vedlejší věta pravděpodobně není vztažná, nýbrž předmětná (17). Opravu lze také provést uvnitř věty u obratů jako *ten který* (18). Zde někdy parser označuje hned dva chybné subjekty (jak slovo *ten*, tak *který*).

Ukazovací zájmeno *ten* (*to*) je často označeno jako ***Sb*** správně. Opravu, tj. změnu závislosti na následující substantivum a změnu funkce na ***Atr***, lze provést, jen pokud se toto zájmeno shoduje s následujícím substantivem a jsou vyloučeny jiné možnosti řešení: sloveso není tranzitivní (19) nebo na něm závisí jiný předmět v akuzativu (20); na slovesu *být* závisí jiné slovo označené jako ***Pnom*** (21).

(15) *Soudní spor/**Sb** , který začal před deseti lety , vyšel zatím Prahu 8/**Sb**/**Atr**/**-3**/**-1** na téměř 31 tisíc korun* .

(16) *epicentrem zemětřesení , jehož/**Sb**/**Atr**/**4**/**2** tektonické vlny/**Sb** dodnes zmítají světovou ekonomikou* .

(17) *Především nevíme , které/**Sb**/**Atr**/**5**/**1** kompetence/**Sb** na nás budou převedeny* .

(18) *na jejímž základě je ten/**Sb**/**Atr**/**3**/**2** který/**Sb**/**Atr**/**2**/**1** jedinec/**Sb** přiřazován do předem známé skupiny* ,

(19) *Mosca s úlevou pozoroval , že to/**Sb**/**Atr**/**2**/**1** dítě/**Sb** vypadá skoro přesně jako jeho* .(20) *bude to naprostá pravda , jenom to/**Sb**/**Atr**/**2**/**1** město/**Sb** je jiné/**Pnom*** ,

(21) *O to víc mě/**Obj** to/**Sb**/**Atr**/**3**/**1** jeho sobotní rozhodnutí/**Sb** šokovalo* .

2.3.2.7 Dva subjekty závislé na jednom slovese, jeden subjekt přes hranici klauzí

Relativně málo se vyskytují struktury, kde jsou dva subjekty závislé na témž slovese, přičemž závislost jednoho subjektu jde přes hranici klauzí, přestože v rámci jeho vlastní klauze sloveso je. Pokud je to možné (vzhledem ke shodě, jinému subjektu blízkého slovesa aj.), převěsí se subjekt na nejbližší sloveso (22), (23) a (24). Tento obecnější typ chyby systematicky opravuje pravidlo 2.2.1, ale za určitých okolností nemusí být předřazeno pravidlu pro dva subjekty, takže i zde se musí takové struktury samostatně opravovat.

(22) *Bývala to posluchárna/**Sb**/**6**/**-2** , ale zbytek/**Sb** budovy se proměnil v ruiny .*

(23) *Tři miny/**Sb**/**7**/**1** byly neoznačené a 40/**Sb** jich bylo označeno jako cvičné .*

(24) *Proud už zase nejde a voda neteče a zběsilci/**Sb**/**-2**/**6** s ostřelovacími puškami se dnes trefují do lidí*

Pravidlo pro opravu dvou subjektů závislých na jednom slovese						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	47	45	628	273	0	993
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	79 %		19 %		2 %	

2.3.3 Pravidlo pro opravu podmětu v bezpředložkovém akuzativu

Protože i trénovací data obsahují jisté procento chyb, mohou se ve výsledcích parseru objevit kombinace morfologických značek a syntaktických funkcí, které jsou z lingvistického pohledu zcela vyloučené, například funkce podmět u syntaktického substantiva v bezpředložkovém akuzativu. Někdy je skutečně chybná syntaktická funkce (1), ale v poměrně velké části těchto případů je syntaktická struktura (závislosti) i syntaktická funkce určena správně, chybná je morfologická značka. Obvykle se chybně disambiguuje akuzativ místo nominativu (2), výjimečně by byl namístě jiný pád (3).

(1) *Ale nevěřil , že jsou přesně to , co/**PQ--4**/**Sb** o sobě říkají .*

(2) *porostem zežloutlých listnatých stromů , se prudce a neočekávaně hnal silný vítr/**NNIS4**/**Sb** .*

(3) *Podobné prezentace/**NNFP4**/**Sb** se jí dostalo od konkurenční stanice Capital FM*

Pravidlo se zavolá na všechny větné členy přímo závislé na slovese, které mají syntaktickou funkci ***Sb*** a zároveň jsou (podle morfologické značky) v akuzativu. Velkou většinu případů lze vyřešit změnou pádu z akuzativu na nominativ (vykazuje-li slovo homonymii akuzativ–nominativ), samozřejmě za přísně stanovených podmínek, jiné špatně disambiguované pádové homonymie jsou mnohem méně časté. Ve zbytku případů pravidlo mění syntaktickou funkci: na předmět nebo příslovečné určení. Bez možnosti sledovat širší kontext a s chybnou disambiguací jako vstupem (spolehlivé řešení homonymie nominativ–akuzativ patří k nejsložitějším úkolům při disambiguaci češtiny) nemůže být pravidlo stoprocentně spolehlivé, ale v celkovém součtu přispívá k celkové kvalitě značkování (morfologického i syntaktického).

Pravidlo nejprve ověřuje, zda je na řídícím slovese závislý jiný subjekt (včetně nevyjádřeného subjektu první nebo druhé osoby) nebo objekt v akuzativu, dále zda je zkoumaný subjekt homonymní s nominativem a zda by se nominativ shodoval se slovesem v čísle (popř. v čísle a rodě).

Jestliže je akuzativní subjekt homonymní s nominativem (stejně jako případné shodné přívlastky závislé na akuzativním subjektu), sloveso se s nominativem může shodovat a nemá jiný subjekt, změní pravidlo morfologickou značku subjektu z akuzativ na nominativ (4) a (5). V opačném případě pravidlo nejprve ověří, zda akuzativní subjekt nepatří do skupiny typicky časových durativních adverbiálií (*hodina*, *chvíle*); pokud ano, změní syntaktickou funkci na ***Adv*** (6).

Jinak ověří, zda sloveso nemá jiný objekt v akuzativu, zda je sloveso tranzitivní a není v pasivním tvaru; jestliže jsou všechny podmínky splněny, změní syntaktickou funkci na ***Obj*** (7). Není-li možná ani tato změna, neprovede se oprava vůbec. To, že na slovese není závislý jiný subjekt a že se sloveso může shodovat s (nominativem) subjektu, není samozřejmě zárukou, že je namístě změna morfologické značky, a ne změna syntaktické funkce, opravný algoritmus tedy není zcela spolehlivý (8) a (9): chybná oprava je vyznačena oranžovou barvou.

(4) *Povstání/**NNNS4**/**NNNS1**/**Sb** skončilo z vojenského hlediska katastrofou , ale mělo nemalý politický význam .*

(5) *Pro Rusy všechno/**PLNS4**/**PLNS1**/**Sb** skončilo špatně .*

(6) *Každý večer/**NNIS4**/**Sb**/**Adv** nám připravila něco k jídlu a čekala v posteli*

(7) *Pokud by ale návrh/**NNIS4**/**Sb**/**Obj** nedal sám majitel , udělali bychom to my*

(8) *A je nám prd/**NNNS4**/**NNNS1**/**Sb** platný .*

(9) *Přivolali koně/**NNMP4**/**NNMP1**/**Sb** , rychle z nich sundali cestovní koše*

Pravidlo pro opravu podmětu v bezpředložkovém akuzativu						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	0	13	573	0	301	887
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	68 %		24 %		8 %	

2.3.4 Pravidlo pro opravu koordinace nekompatibilních syntaktických funkcí

Když je větný člen závislý na koordinaci a tato koordinace na dalším větném členu, je pro parser určení syntaktické funkce závislého stromu obtížnější, protože se mezi skutečným řídícím členem a podřízeným větným členem nachází jedna úroveň navíc. Zároveň parser nedokáže ověřit sousední uzly. Ve výstupu parseru se tak často objevují koordinované větné členy, které mají různé, nekompatibilní syntaktické funkce (1) a (2).

Oprava se zaměřuje pouze na koordinovaná syntaktická substantiva, jež se shodují v pádu (tj. koordinace je pravděpodobně v pořádku, stačí opravit syntaktickou funkci). Opravu koordinace syntaktických substantiv v různých pádech bude nutné do opravného programu doplnit, ale vzhledem k ne zcela spolehlivé disambiguaci a k tomu, že by pravděpodobně bylo nutné zasahovat do celé závislostní struktury, je to úkol mnohem složitější a nebyl dosud vyřešen.

(1) *V říjnu/**Atr** a/**Coord** listopadu/**Adv** Francouzi zlikvidovali výběžek fronty ve středu německé linie.*

(2) *starým Brunem , který ho naučil lovit oštěpem/**Pnom** a/**Coord** prakem/**Atr** i žít způsobem lidí klanu .*

Pravidlo se spustí, když hlavní program narazí na dvě koordinovaná syntaktická substantiva s rozdílnou syntaktickou funkcí. Pravidlo v tom případě ověří ostatní členy koordinace. Jestliže je mezi členy koordinace nějaký nesourodý prvek (sloveso, jiný pád aj.), mělo by tuto zřejmě chybnou koordinaci opravit jiné pravidlo: pravidlo se neuplatní.

Jsou-li členy koordinace pouze syntaktická substantiva ve stejném pádě, ověří se, zda řídící uzel koordinace, přiřazený parserem, je nejlepší možný, případně bude vyhledán vhodnější na základě valence sloves, adjektiv, rekce substantiv aj. (3). Podle řídícího větného členu (řídící člen koordinace většinou zůstává nezměněn) pak pravidlo zvolí vhodnou syntaktickou funkci (4). Je-li řídícím uzlem předložka, pak zvolí funkci podle řídícího větného členu předložky (5) a (6).

(3) *a sepsal spisek Průvodce po/**-3**/**-1** Písku/**Atr**/**1** a okolí/**Adv**/**Atr**/**-1***

(4) *Jeden má ženu/**Obj** a dům/**Atr**/**Obj** a všecko , nač si vzpomene , a druhý nemá nic .*

(5) *Také toužím po teplé vodě/**Adv**/**Obj** a splachovacím záchodě/**Atr**/**Obj** .*

(6) *které se objevily v novinách/**Atr**/**Adv** a časopisech/**Adv** .*

Pravidlo pro opravu koordinace nekompatibilních syntaktických funkcí						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	24	1	576	0	0	601
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	82 %		12 %		6 %	

2.3.5 Pravidlo pro opravu přímého předmětu závislého na netranzitivním slovese

Předmět v bezpředložkovém akuzativu může být správně závislý pouze na tranzitivním slovese v činném rodu. Parser však struktury, které tomuto pravidlu odporují, opakovaně vytváří, zčásti proto, že mu trénovací data neposkytují dostatečné údaje o valenci méně frekventovaných sloves. Někdy dává chybně přednost závislosti akuzativního předmětu na pomocném slovese (1). Obtížně rozlišuje mezi valenčními předměty a časovými či měrovými určeními v bezpředložkovém akuzativu, zvláště v případě kvantifikace (2). Situaci dále komplikuje chybná disambiguace pádové homonymie nominativ–akuzativ (3).

(1) *Proč jsi mě/**Obj**/**-1** nikdy nevyslechla , proč ses o to ani jednou nepokusila .*

(2) *Podlitiny přetrvávají asi deset/**Obj** dní .*

(3) *Tady už z něj netryskal proud/**NNIS4**/**Obj** pražských legend a pařížských historek ,*

Opravný program volá toto pravidlo, když narazí na netranzitivní sloveso (tuto vlastnost ověřuje v seznamu sloves s jejich valencemi) rozvité předmětem v bezpředložkovém akuzativu (předmět závisí přímo na slovesu). Pokud zkoumaný předmět závisí na slovese *být*, jež má funkci pomocného slovesa (nemusí být nutně označeno ***AuxV***, stačí, když např. tvar přítomného času slovesa *být* rozvíjí minulé nebo trpné příčestí), bude převěšen na plnovýznamové sloveso. Je-li toto sloveso tranzitivní, změna je dokončena (4), jinak pokračují změny jako pro jiná, plnovýznamová slovesa.

Jestliže zkoumaný předmět patří mezi substantiva vyjadřující čas nebo jde o číslovku (základní číslovku *pět* a více v akuzativu, popř. číslovku neurčitou), která kvantifikuje taková substantiva, bude opravena syntaktická funkce z ***Obj*** na ***Adv*** (5) a (6). Toto řešení však není zcela spolehlivé, v

některých spojeních opravuje struktury špatně (7); bylo by vhodné ještě rozlišit, u kterých sloves má časové vyjádření obvykle funkci subjektu, a ne příslovečného určení (např. *ubíhat*, *plynout*), popř. přísně vyžadovat kvantifikaci časových údajů u plurálu.

Je-li tvar předmětu pádově homonymní (nominativ–akuzativ), sloveso nemá jiný subjekt, je ve třetí osobě a může se shodovat s potenciálním nominativem u předmětu, změní pravidlo morfologickou značku (akuzativ na nominativ) a syntaktickou funkci slova z ***Obj*** na ***Sb*** (8). Ani tato změna není zcela spolehlivá, zvlášť ve vztahu ke slovu *to*, jak je zřejmé z příkladu (9), kde byla chybně opravena syntaktická funkce slova *to* místo jeho závislosti (chybná oprava je vyznačena oranžovou barvou), čímž byla zablokována správná oprava následujícího slova. Pro správné určení závislostí, funkcí, pádu či dokonce slovního druhu tvaru *to* by se však mělo vytvořit samostatné pravidlo; tato problematika je velmi složitá.

(4) *kdo psal ten vzkaz , který/**Obj**/**1**/**2** jsi našel na stole ,*

(5) *Děj , který trval ve skutečnosti tři vteřiny/**Obj**/**Adv** , proběhne na projekční stěně*

(6) *Lord seděl několik/**Obj**/**Adv** vteřin/**Atr**/**-1** beze slova a sledoval rybky*

(7) *Vteřiny/**Obj**/**Adv** ubíhaly , pět , možná deset , možná dvacet .*

(8) *jednoho dne u mě zazvonil telefon/**NNIS4**/**NNIS1**/**Obj**/**Sb** a naše dispečerka nasadila*

(9) *ale kdepak , to/**PDNS4**/**PDNS1**/**Obj**/**Sb**/**2** děťátko/**NNNS4**/**Obj** hovořilo ke mně !*

Pravidlo pro opravu přímého předmětu závislého na netranzitivním slovese						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	13	0	109	188	0	310
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	85 %		15 %		0 %	

2.3.6 Pravidlo pro opravu předmětu závislého na modálním či fázovém slovese

Na některých modálních a fázových slovesech (*chtít*, *začít*) mohou být závislá syntaktická substantiva s funkcí předmětu. Pokud je však na modálním či fázovém slovesu závislé plnovýznamové sloveso v infinitivu, obvykle je předmět závislý na něm. Pro parser je takové rozhodování obtížné, parser často raději zvolí bližší sloveso nebo sloveso v určitém tvaru. Vznikají tak případy, kdy je syntaktické substantivum v bezpředložkovém akuzativu závislé na modálním či fázovém slovese, které má infinitivní předmět (1), nebo na modálním či fázovém slovese místo na slovese plnovýznamovém závisí předmět v předložkové frázi, plnovýznamové sloveso je přitom valenční s odpovídající valencí (2).

(1) *Slova/**Obj**/**3** ze sebe musel přímo páčit .*

(2) *To je úroveň , o/**3** které/**Obj** si může většina Evropy od dob ropných krizí v sedmdesátých letech nechat jen zdát .*

Pravidlo se spustí z hlavního programu, když je na modálním či fázovém slovese zároveň závislý předmět v infinitivu i syntaktické substantivum s funkcí ***Obj***, a to buď v bezpředložkovém akuzativu, nebo v předložkové frázi. Pravidlo se dělí do dvou větví, jedna opravuje závislost či značkování předmětu v akuzativu, druhá se zaměřuje na předložkové fráze.

V případě s předmětem v bezpředložkovém akuzativu je náprava chyby složitější. Stejně jako u výše popsané chyby se dvěma subjekty závislými na jednom slovese je častou příčinou této chyby nesprávná disambiguace pádově homonymních slov, zvláště homonymie nominativ–akuzativ.

Jestliže modální či fázové sloveso nemá jiný podmět, je ve třetí osobě a potenciální tvar nominativu u zkoumaného předmětu se shoduje se slovesem v rodě a čísle, je pravděpodobnější, že je třeba opravit morfologickou značku a funkci předmětu spíše než měnit závislosti ve větě (3). Má-li plnovýznamové sloveso v infinitivu již jiný předmět v akuzativu, je toto řešení víceméně jediné možné (4). Takováto oprava opět není zcela spolehlivá, identifikovaná chyba může být způsobena jinou chybou na jiném místě (5): chybná oprava vyznačena oranžovou barvou.

Nelze-li změnu z nějakého důvodu provést (např. proto, že předmět nevykazuje homonymii ak.–nom., na slovese již je závislý jiný subjekt aj.), převěsí pravidlo předmět na plnovýznamové sloveso, ovšem pouze pokud na druhém slovese není závislý jiný přímý předmět a sloveso je tranzitivní (6).

(3) *Jedná se o zcela novou trať , po které by vlaky/**NNIP4**/**NNIP1**/**Obj**/**Sb**/**1** mohly jezdit rychlostí kolem 300 kilometrů za hodinu ,*

(4) *Kvůli nepravidelnému rozložení váhy nemohou nohy/**NNFP4**/**NNFP1**/**Obj**/**Sb**/**-2** zcela absorbovat otřesy/**NNIP4**/**Obj**/**-1** vyvolané chůzí .*

(5) *zda vyhlášenému hlavičkáři nemohly potíže/**NNFP4**/**NNFP1**/**Obj**/**Sb**/**-1** způsobit například právě údery/**NNIP4**/**Obj**/**-3** míče .*

(6) *Musela jsem si Prahu/**NNFS4**/**Obj**/**-3**/**4** teprve zase postupně ohmatat .*

Ve druhé větvi pravidla je oprava chyby jednodušší: je-li na jednom modálním či fázovém slovese závislá předložková fráze s funkcí předmětu a zároveň sloveso v infinitivu s odpovídající valencí (předložka a pád), změní opravný program závislost předložkové fráze z modálního slovesa na valenční sloveso v infinitivu (7). V některých specifických případech je nutné převěsit předmět na infinitiv slovesa, který je v závislostní struktuře ještě o stupeň níže (8).

(7) *Stát se o/**4**/**3** mě/**Obj** bude muset postarat - koneckonců jsem přece nežádal o to , abych se narodil .*

(8) *že by si o/**2**/**8** tom/**Obj** mohl pravidelný konzument hororů nechat jenom zdát .*

Pravidlo pro opravu předmětu závislého na modálním či fázovém slovese						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	95	0	3	16	0	114
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	87 %		4 %		9 %	

2.4 Pravidla pro opravu závislostí a syntaktických funkcí předložkových frází

Správně určit závislosti a syntaktické funkce v předložkových frázích (především přímo závislosti předložek a větných členů na nich přímo závislých) je pro parser obtížnější než určení závislostí syntaktických substantiv v bezpředložkových pádech, mimo jiné i proto, že kombinací předložek a pádů je mnohem více než prostých pádů a v trénovacích datech se parser s jednotlivými kombinacemi předložky, pádu a syntaktické funkce setká méně často.

Obtížné je i určení (formální) závislosti. Na předložce obvykle závisí syntaktické substantivum ve stejném pádu, jako vyžaduje předložka, nebo na ní závisí koordinační spojka, jež koordinuje syntaktická substantiva, která jsou ve stejném pádu, jako vyžaduje předložka. V případě elipsy substantiva však může být na předložce závislé i syntaktické adjektivum. Parser navíc nedokáže spolehlivě rozlišit syntaktická substantiva a adjektiva. Chyby týkající se předložkových frází jsou velmi rozmanité, zaměřuje se na ně nejvíc pravidel v celém opravném programu.

2.4.1 Pravidlo pro opravu syntaktických funkcí větných členů závislých na předložce

Velmi častou chybou použitého parseru jsou chybně přiřazené syntaktické funkce větných členů závislých na předložce, popřípadě chybně určená závislost celé předložkové fráze. Syntaktickou funkci nese člen závislý na předložce, mezi řídícím uzlem a závislým členem jsou tak dvě úrovně, což parseru určení funkce ztěžuje. Částečným řešením je poměrně jednoduché pravidlo na opravu takto chybných struktur. Následující tři příklady ukazují typy chyb, na něž se pravidlo zaměřuje: předložkové fráze závislé na substantivu s označením ***Adv*** nebo ***Obj*** (1) a (2) a předložkové fráze závislé na slovese či adjektivu s označením ***Atr*** (3).

(1) *Pro vstup do/-**1** války/**Adv** bylo takové vysvětlení podle jeho názoru nezbytné*

(2) *Ve starých domech není o/**2** výklenky/**Obj** nouze .*

(3) *Němci dosahují průlomu u/-**2** Gorlice/**Atr** a vytlačují Rusy z Karpat .*

Pravidlo se volá v případě, že na substantivu, adjektivu, slovese nebo vybraných zájmenech a číslovkách je přes předložku závislý větný člen, který nemá odpovídající syntaktickou funkci. Na substantivu musí být závislý větný člen s funkcí ***Atr***. Je-li větný člen v genitivu plurálu závislý přes předložku *z* na adjektivu, číslovce či zájmenu typu *jeden*, *první*, *další*, *některý*, *každý* (jež jsou v daném kontextu syntaktickými substantivy), musí mít také funkci ***Atr***. Větný člen závislý na slovese či adjektivu musí mít funkci ***Obj*** nebo ***Adv*** (subjekty v předložkové frázi řešíme samostatným pravidlem; jmenné části verbonominálních predikátů v předložkových frázích se v PDT označují jako ***Adv***).

Pravidlo opravuje funkci tak, aby odpovídala slovnímu druhu větného členu, na němž je předložka závislá, pokud nemá zvláštní důvod ke změně závislosti předložky na jiný větný člen ve větě. K posouzení takové změny se pravidlo rozděluje na dvě větve: předložka závislá na syntaktickém substantivu a předložka závislá na slovese či přídavném jménu.

2.4.1.1 Chybné syntaktické funkce závislé přes předložku na (syntaktickém) substantivu

Je-li podstatné jméno předcházející předložce „valenční“ (předložková fráze s danou předložkou a pádem obvykle závisí na něm: *řeč o 6*, *debata o 6*, *kniha o 6*...), pravidlo změní syntaktickou funkci na ***Atr*** (4). Stejně tak pro větné členy závislé na syntaktickém substantivu typu *první*, *každý* přes předložku *z*.

Pokud řídící substantivum není valenční, zjistí pravidlo, zda se v klauzi nenachází sloveso s odpovídající předložkovou valencí (*platit za* 4); pak změní závislost předložky, funkce bude ***Obj*** (5).

V opačném případě pravidlo ověří, zda se v klauzi nachází sloveso, na němž často závisí předložková fráze s daným pádem a předložkou s adverbiální funkcí (*schovat za 4*); pak změní závislost předložka, funkce bude ***Adv***.

Dále se změní závislost předložky, pokud se v klauzi nachází sloveso a předložka závisí na propriu – propria jsou jen zřídka rozvíjena předložkovými frázemi (6), nebo předložková fráze patří k typicky adverbiálním určením – zvl. časovým nebo místním (7). Tato poslední možnost ovšem není zcela spolehlivá (8). Ve zbývajících případech zůstane závislost zachována a syntaktická funkce bude opravena na ***Atr*** (9).

(4) *před sebou však viděl velkou naději na/-1 medaili/**Obj**/**Atr***

(5) *V sále na/-1/3 šťastný pár/**Adv**/**Obj**/-2 čekaly tři stovky pozvaných hostů ,*

(6) *Český Telecom v/-1/2 současnosti/**Adv** rozmisťuje telefonní automaty s cílem zvýšit jejich využívání .*

(7) *která se narodila císařským řezem ve/-3/-1 čtvrtek/**Adv** v Ústavu péče o matku a dítě v Podolí*

(8) *celého tohoto pokoření , v němž vidí znovu pokoření z/-1/-3 roku/**Adv** 1844*

(9) *Bože na/-1 nebesích/**Adv**/**Atr** , " bědovala Maja-Lisa .*

2.4.1.2 Chybné syntaktické funkce závislé přes předložku na slovese nebo adjektivu

Je-li sloveso nebo adjektivum, na němž je předložka závislá, valenční (*nutit k 3, přimět k 3; šetrný k 3, nutný k 3*), pravidlo změní syntaktickou funkci na ***Obj*** (10) a (11). Pokud řídící sloveso patří do seznamu sloves, na nichž jsou často závislé předložkové fráze s danou předložkou a pádem s adverbiální funkcí, nebo je-li sama předložková fráze typicky okolnostní (12), změní se funkce na ***Adv***.

Jinak pravidlo ověří, zda těsně před předložkou nestojí „valenční“ (rekční) substantivum s odpovídající předložkovou vazbou. Jestliže ano, změní se závislost předložky na toto substantivum a funkce na ***Atr*** (pokud již původní funkce nebyla ***Atr***). Ve zbývajících případech zůstane závislost zachována a syntaktická funkce bude opravena na ***Adv***; ne vždy je to ale správně, viz (13): chybná oprava je zvýrazněna oranžovou barvou.

(10) *V praxi však – a kůň o/2 tom/**Atr**/**Obj** věděl své – takových čtyřiadvacet hodin*

(11) *Rakousko-Uhersko označilo vrahy za/-2 loutky/**Atr**/**Obj** srbské zpravodajské služby*

(12) *Ve Francii v/5 prvních třech letech/**Atr**/**Adv** války kráčela diktatura a koalice většinou ruku v ruce .*

(13) *co se dělo s lidmi z/-3 jejich okolí/**Atr**/**Adv***

Pravidlo pro opravu syntaktických funkcí větných členů závislých na předložce						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	699	13	7161	0	0	7873
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	70 %		29 %		1 %	

2.4.2 Oprava chybného určení syntaktické funkce substantiva v předložkové frázi závislé na slovese

I když parser správně určí závislost předložkové fráze, označení syntaktické funkce bývá problematické. Parseru scházejí rozsáhlá data o předložkových valencích sloves, nadto není hranice mezi příslovečným určením a předmětem v případě předložkových frází zcela ostrá a ani v trénovacích datech PDT se nehodnotí všechny případy zcela jednotně. Přesto lze s využitím dat z rozsáhlých korpusů poměrně spolehlivě mezi předložkovými frázemi závislými na slovesech rozlišovat a vybrat jim vhodné syntaktické funkce.

To, že problémy parseru vycházejí primárně z relativně malých trénovacích dat, dokládá skutečnost, že u často se vyskytujících valenčních sloves i u frekventovaných příslovečných výrazů jsou chyby výjimečné (substantiva v lokálu v předložkové frázi s předložkou *o* závislé na slovese *mluvit* jsou v 99 % označené jako předmět; spojení *o půlnoci* závislé na slovese je v 97 % příslovečné určení), protože se dostatečně často vyskytovaly v trénovacích datech, zatímco u méně častých valenčních sloves a příslovečných určení je procento chyb mnohem vyšší, například substantiva v lokálu v předložkové frázi s předložkou *po* závislé na slovese *dychtit* jsou v 17 % označena chybně jako příslovečná určení (1); *o Letnicích* je ve 36 % označeno chybně jako předmět (2).

(1) *Jsou lidé , kteří dychtí po tom/**Adv** , udělat ze světa jeden jediný kriminál*

(2) *Mimoto jste mi říkal , že býváte o letnicích/**Obj** v Praze u Mistra Bílka .*

Jedna větev pravidla se spouští na předložkové fráze označené ***Adv***, jež jsou závislé na slovese, které má valenci s danou předložkou a pádem. Druhou větev volá hlavní program na předložkové fráze označené ***Obj***, které jsou závislé na slovese, jež nemá odpovídající valenci.

Valence se ověřují v seznamu cca 2000 sloves získaného z korpusů SYN2005 a SYN2010. Seznam zahrnuje jednak slovesa s předložkovou valencí (např. *toužit po 6*), jednak slovesa, jež jsou typicky rozvíjena předložkovými frázemi s okolnostním významem (nevalenční, např. *schovat se za 4*). Seznam odlišuje slovesa, která mají určitou valenci pouze jako zvratná (*pustit se do 2*, *vyslovit se k 3*). Kromě toho jsme sestavili seznamy typických adverbiálních předložkových frází (cca 1300 kombinací lemmat substantiv s předložkou a pádem, např. *o víkendu*, *o Vánocích*; *za chvíli*, *za týden*). Tyto seznamy slouží jako základ pro opravné pravidlo (používají se však i v mnoha dalších pravidlech).

Je-li jako příslovečné určení (***Adv***) označeno substantivum v předložkové frázi závislé na slovese, jež má odpovídající valenci, a předložková fráze nepatří mezi typická příslovečná určení s daným

pádem a předložkou, bude syntaktická funkce změněna z ***Adv*** na ***Obj*** (3) a (4). Typická příslovečná fráze změněna nebude (5), přestože někdy tato opatrnost není namístě (6).

(3) *Fischer by na toto vysvětlení/**Adv**/**Obj** pohlížel určitě skepticky .*
(4) *Orlin Grabbe zjistil , že se na ní/**Adv**/**Obj** podílejí dvě skupiny*
(5) *Daisy a Tom na sebe na okamžik/**Adv** mlčky pohlédli .*
(6) *Celý Šumperk čeká na chvíli/**Adv** , kdy se domů vrátí nový hrdina*

Ve druhé větvi směřuje oprava opačným směrem: patří-li řídící sloveso předložkové fráze do seznamu sloves, která nemají danou valenci, ale naopak jsou obvykle rozvíjena okolnostními určeními s danou předložkou a pádem, bude funkce syntaktického substantiva v předložkové frázi změněna z ***Obj*** na ***Adv*** (7). Stejná oprava funkce se provede, jestliže řídící sloveso nepatří do žádného seznamu (tj. není u něj evidována daná valence) a předložková fráze patří mezi typicky adverbiální (8).

(7) *Jedu na chalupu/**Obj**/**Adv** , je , mmm , čtvrtek večer , 8.47 .*
(8) *Muž se s očividnou úlevou/**Obj**/**Adv** protahuje dovnitř .*

Stojí-li před předložkovou frází, která je závislá na nevalenčním slovese, substantivum s rekcí odpovídající předložce a pádu, převěsí pravidlo předložku na substantivum a změní funkci substantiva v předložkové frázi (9). Je-li předložková fráze závislá na modálním či fázovém slovese, na němž je závislé také sloveso v infinitivu, které má odpovídající valenci, převěsí pravidlo předložkovou frázi na sloveso v infinitivu (10). Obě změny by za normálních okolností měla provést jiná pravidla, ale protože pořadí prováděných oprav může být ovlivněno mnoha faktory, je nutné, aby byly možnosti takových změn zahrnuty i do tohoto pravidla.

(9) *V půlce koncertu pak na scénu zavítal hráč na/**-2**/**-1** bicí/**Obj**/**Atr** , což bylo skutečně*
(10) *Nechci na/**-1**/**3** to/**Obj** ani pomyslet .*

Výsledky testování pravidla ukazují, že je ještě nutné doplnit do opravného programu změnu morfologické značky místo změny funkce, protože některé chyby jsou způsobeny chybnou disambiguací pádu: u pádově homonymního substantiva s pádově homonymní předložkou (např. akuzativ/lokál) závislého na slovese, které má valenci s danou předložkou, ale s jiným pádem, je třeba změnit pád v morfologické značce (11), dosud však tato možnost implementována nebyla.

(11) *Na zajištění/**NNNS6**/**Adv** tahače lany čekali i hasiči*

Oprava chybného určení syntaktické funkce substantiva v předložkové frázi závislé na slovese						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	65	5	2868	0	0	2938
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	87 %		5 %		8 %	

2.4.3 Pravidlo pro určení závislosti předložkových frází na rekčních substantivech

Určit závislost předložkové fráze je často obtížné i pro trénovaného anotátora, natož pro parser, který se může opřít jen o omezená trénovací data, a ne o jazykové kompetence rodilého mluvčího. Volba řídícího uzlu předložkové fráze je často ovlivněna nejen řídícím uzlem (jeho slovním druhem, valencí, rekcí aj.) a lemmatem a pádem předložky, ale také významem celé předložkové fráze (zjednodušeně lemmatem syntaktického substantiva v předložkové frázi). Parser tak v určení závislosti předložkových frází chybuje častěji než u nepředložkových pádů syntaktických substantiv (1). Pravidlo vyhledává rekční substantiva (substantiva, jež jsou často rozvíjena určitou typickou kombinací předložky a pádu), po nichž následuje předložka s odpovídajícím lemmatem a pádem. Není-li na substantivu taková předložková fráze závislá, ve vhodných případech závislost předložky mění. Mezi chybnými strukturami z výstupů parseru se často objevují struktury, kde syntaktickou funkci substantiva (***Atr***) není nutné měnit, přestože předložka není závislá na substantivu. Je to tedy chyba, kterou by případně mohlo řešit i pravidlo pro opravu syntaktických funkcí větných členů závislých na předložce (2).

(1) *Pravděpodobně však špatná manipulace s/**2** nákladem/**Adv** způsobila , že vozík*

(2) *Arabské povstání vyrostlo z touhy po/**-3** nezávislosti/**Atr** na Turecku , která byla*

Pravidlo se zavolá, když těsně po substantivu, které patří do seznamu rekčních substantiv (např. *touha po 6*), následuje odpovídající předložka (např. *po* s lokálem), přičemž tato předložka není závislá na rekčním substantivu. Pravidlo nezasáhne do struktury, je-li předložka závislá na valenčním slovese či adjektivu (jež také vyžadují tuto předložku). Pravidlo také neprovede žádnou změnu, když předložka se substantivem patří do seznamu předložkových frází s časovým či lokálním významem, např. *o Vánocích* (3).

Jestliže se žádná taková překážka neobjeví, pravidlo převěsí předložku na rekční substantivum

(4), případně také změní syntaktickou funkci slova závislého na předložce (5). Když je předložka závislá na následující koordinační spojce, ponechá pravidlo ve většině případů závislost nezměněnou, pokud není v koordinaci jednoznačně chyba, např. když koordinační spojka zároveň koordinuje slovesa i předložky (6), ani pak ale nedokáže opravit strukturu celou.

(3) *Návrhy sledují recepty používané s úspěchem v/**-3** zemích/**Adv** třetího světa*

(4) *Řekl jsem mu , že jsem víru/**2**/**-1** v Boha ztratil .*

(5) *Tuhle cestu do/**6**/**-1** Česka/**Adv**/**Atr** si Bulhaři za rámeček nedají .*

(6) *Když obři zmizeli , otevřel se vchod do/**2**/**-1** skály/**Atr** a/**-3** v něm stál/**-3** Henke .*

Pravidlo pro určení závislosti předložkových frází na rekčních substantivech						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	218	0	17	0	0	235
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	85 %		6 %		9 %	

2.4.4 Předložkové fráze jako přívlastky substantiv bez odpovídající rekce

Volba správného řídícího členu předložkové fráze je často problematická. Ve větě, kde se nachází sloveso s odpovídající předložkovou valencí je však pravděpodobné, že řídícím členem předložkové fráze má být sloveso. Parser však nemá k dispozici dostatečná data o valenci méně frekventovaných sloves ani o rekčních substantivech, nedokáže proto spolehlivě rozlišit mezi přívlastkem substantiva a předložkovým předmětem slovesa a chybně určuje závislost předložkové fráze na substantivu (1) a (2).

(1) *Je tedy třeba , aby rodiče nebo lékař s/**-1** učitelem/**Atr** promluvili*

(2) *skoro každej den koukáme v kostelech na/**-1** svatby/**Atr** .*

Opravný program spustí toto pravidlo, když identifikuje předložkovou frázi závislou na substantivu, které nemá odpovídající rekci (není v seznamu rekčních substantiv nebo v daném seznamu nemá přiřazenu danou kombinaci předložky a pádu). To samo o sobě neznamená chybu. Pravidlo se pokusí najít sloveso ve stejné klauzi. Najde-li sloveso, které má předložkovou valenci s danou předložkou a pádem (*dívat se po 6*, *pohrávat si s 7*), změní závislost předložkové fráze na toto sloveso (3). Výjimkou je případ, kdy je na valenčním slovesu závislá jiná předložková fráze se stejnou předložkou a substantivem s funkcí ***Obj*** , tj. valenční (4). Pravidlo může také využít infinitivní předměty modálních a fázových sloves jako řídící uzly předložkových frází, mají-li správnou předložkovou valenci (5).

(3) *Dívala jsem se z okna tramvaje po/-1/-6 domovních číslech/Atr/Obj ,*

(4) *Dál si pohrával s/-1 kšiltovkou/Obj s/-1 koženým štítkem/Atr .*

(5) *Matka o/-1/4 morčeti/Atr/Obj nechtěla ani slyšet .*

Předložkové fráze jako přívlastky substantiv bez odpovídající rekce						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	160	0	159	0	0	319
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	76 %		9 %		15 %	

2.4.5 Pravidlo pro opravu koordinací závislých na předložkách

Určení závislosti koordinace následující po předložce a jejích členů je pro parser často obtížné, parser nedokáže spolehlivě rozlišit (například na základě slovního druhu a pádu členů koordinace), zda má být koordinace závislá na předložce nebo na jiném větném členu. Parser také často přiřazuje předložce více samostatných závislých uzlů. V důsledku toho je pak někdy na jedné předložce závislé jak následující syntaktické substantivum, tak koordinace. Někdy by mělo být součástí koordinace i první syntaktické substantivum po předložce (1), jindy koordinace na předložce závislá být nemá (2). Často je koordinace také závislá na chybné (dřívější) předložce (3). Na nespolehlivé určení závislostí v případě koordinační spojky po předložce naráželo i předchozí pravidlo pro předložky s neuspokojenou rekcí.

(1) *v povinnostech a věcech cti , ve víře v Boha/-1 a/-2 jeho zákony/-2 !*

(2) *Nejčastěji se horší na podzim/-1 a/-2 na/7 jaře , při účasti pylové alergie i v létě ,*

(3) *Najdeme ji nejen v Řecku , ale i jinde na Balkáně , v Itálii/1 a/-5 jižní Francii/-4 .*

Pravidlo se spustí, kdykoli je na předložce závislá koordinace. Ověří, zda je struktura v pořádku: pokud je na předložce kromě koordinace závislý i jiný větný člen nebo pokud jsou na koordinaci závislé nesourodé prvky, pokusí se pravidlo zvolit vhodný postup opravy.

Je-li na jedné předložce závislé syntaktické substantivum ve správném pádu, které následuje po předložce a stojí před koordinací, ověří pravidlo ostatní členy koordinace a slova mezi tímto syntaktickým substantivem a koordinací. Pokud jsou členy koordinace pouze syntaktická substantiva se stejným pádem jako první syntaktické substantivum a mezi prvním syntaktickým

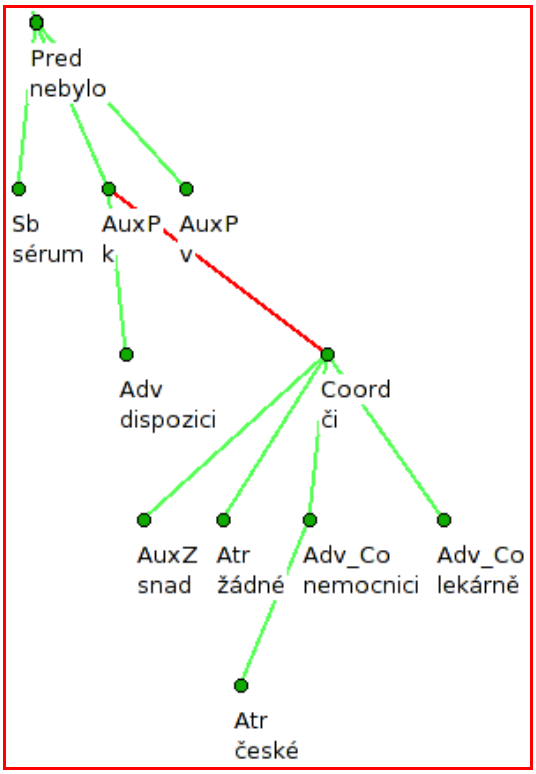
substantivem a koordinací nestojí žádný nevhodný větný člen (sloveso, předložka, podřadicí spojka aj.), převěsí syntaktické substantivum na koordinaci (4).

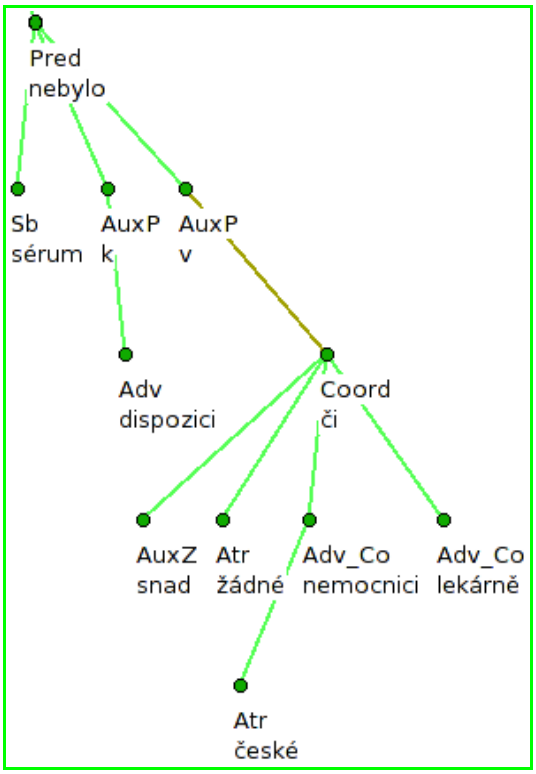
V ostatních případech koordinace nemá být závislá na koordinaci. Pravidlo vyhledá jiný vhodný řídící člen, nejčastěji blízké sloveso, a koordinaci na něj převěsí (5). Jestliže se mezi předložkou a koordinací nachází jiná předložka, jejíž pád odpovídá pádu syntaktických substantiv pod koordinací, převěsí pravidlo koordinaci na tuto předložku (6).

(4) *Znalost cizích právních řádů je pro legislativní záměr/-2/1 a/-3 jeho formulaci/-2 významná z řady důvodů .*

(5) *Kromě toho/-1 vojáky/1 a/-3/3 děla teď potřeboval zcela jinde .*

(6) *sérum nebylo k dispozici snad v žádné české nemocnici či lekárně !*





Pravidlo pro opravu koordinací závislých na předložkách						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	83	4	40	0	0	127
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	70 %		19 %		11 %	

2.4.6 Pravidlo pro opravy chyb u víceslovných předložkových výrazů

Pro víceslovné předložkové výrazy (*v souvislosti s*, *směrem k*, *na základě*) má formalismus PDT na analytické rovině zvláštní označení: části takového výrazu jsou označeny ***AuxP*** (stejně jako prosté předložky) a jsou závislé na poslední části výrazu: *v/**AuxP** souladu/**AuxP** s/**AuxP***. V trénovacích datech je takto označeno 25 trojčlenných obratů, celkem cca 300 výskytů (*v souvislosti s*, *s přihlédnutím k*) a několik desítek dvojčlenných obratů, celkem cca 900 výskytů (*vzhledem k*, *směrem do*, *na základě*, *v oblasti*). Většina obratů má velmi nízkou frekvenci, některé se v trénovacích datech objevují jen jednou. Ve verzi trénovacích dat, která byla použita pro trénování parseru, navíc není v označení víceslovných předložkových výrazů úplná shoda, některé z nich jsou jednou označeny jako víceslovný výraz, podruhé se s nimi zachází jako s běžným spojením předložky a substantiva. Nízká frekvence a nejednotnost v trénovacích datech parseru neumožňují víceslovné předložkové výrazy značkovat správně. Parser chybuje v obou směrech: často neoznačuje ani typické víceslovné předložkové výrazy (1) a (2) a naopak označuje jiná spojení předložky a substantiva (či předložky, substantiva a předložky) jako víceslovné předložkové výrazy (3).

(1) *Nebo ho pod/**AuxP**/**5** vlivem/**Adv**/**-1** takové představy doopravdy zabil ?*

(2) *Vzhledem/**Obj**/**11** k/**AuxP**/**-1** neujasněnosti pravidel publikace právních předpisů*

(3) *letos vůbec poprvé počítá v/**AuxP**/**2** rozpočtu/**AuxP**/**1** s/**AuxP**/**-3** částkou 600 tisíc korun na zajištění*

Hlavní program volá toto pravidlo, pokud narazí na víceslovný předložkový výraz (jednoznačný, sporné případy neopravuje), který není správně označen. Volá je i v opačném případě, když jsou po sobě následující slova označena jako víceslovný předložkový výraz, ale nepatří do seznamu složených předložek ani do seznamu výjimek (sporných případů). Názor autora na to, co by mělo nebo nemělo být značkováno jako víceslovný předložkový výraz, není důležitý, rozhodující je převládající značkování v PDT a seznam nepravých předložek uvedený v anotačním manuálu PDT. Sporné výrazy do tohoto seznamu sice nepatří, ale jsou analogické s jinými výrazy, které tam patří (*směrem k* je v seznamu nepravých předložek, *směrem na* nikoli). U sporných výrazů pravidlo nezasahuje ani v jednom směru.

Všem částem identifikovaných víceslovných předložkových výrazů přiřadí pravidlo funkci ***AuxP***, syntaktickému substantivu, které je na výrazu závislé, bude přiřazena funkce ***Adv*** nebo ***Atr*** podle toho, na jakém slovním druhu je výraz závislý (4) a (5).

U víceslovných předložkových výrazů, které jako takové chybně označil parser, je oprava o něco složitější, je nutné najít vhodné řídící větné členy pro jednotlivé části a vhodné syntaktické funkce pro substantiva ve výrazech. Pravidlo volí řídící členy a funkce podle kontextu, pádu substantiv, valence okolních slov, původního řídícího členu výrazu aj. (6) a (7).

(4) *V/**AuxP**/**7**/**1** případě/**Adv**/**AuxP**/**-1**/**6** nežádoucích účinků nebo pochybností se obraťte na svého lékaře nebo lékárníka .*

(5) *Současně/**Adv**/**AuxP**/**9**/**1** s/**AuxP**/**8** uvažovanými legislativními změnami bude však nutné*

(6) *O/**AuxP**/**1**/**4** zájmu/**AuxP**/**Obj**/**3**/**-1** Newcastlu nic nevím .*

(7) *nastínil jako předehru/**AuxP**/**Adv**/**1**/**-2** ke/**AuxP**/**-3**/**-1** svému putování napříč Texasem*

Pravidlo pro opravy chyb u víceslovných předložkových výrazů						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	28	48	3	0	0	79
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	76 %		13 %		11 %	

2.4.7 Pravidlo pro doplnění neuspokojených rekčních požadavků předložek

Ve výstupu parseru se nezřídka objevují případy, kdy normální česká předložka stojící uvnitř věty nemá uspokojen svůj rekční požadavek (na předložce není závislý žádný vhodný větný člen), přestože po předložce následuje syntaktické substantivum s odpovídajícím pádem. Tato chyba málokdy nastane v jednoduchých, přehledných strukturách, spíše se týká vět s neznámými slovy, s větším počtem předložek, se složitější koordinací nebo s apozicí adjektiv.

Někdy je chyba jen důsledkem jiné, závažnější chyby v konstrukci celého závislostního stromu, kterou by mělo před aplikací tohoto pravidla odstranit jiné, obecnější pravidlo. Tak je tomu například v (1), kde je koordinační spojka (*a*) chybně považována za řídící uzel celé struktury (**#**) místo správné závislosti na předložce (*na*), jíž chybí jakýkoli závislý člen.

Pravidlo se snaží správně naplnit rekční požadavky předložky i za cenu dalších změn struktury. Variabilita chybných struktur je však příliš velká, změna může být zablokována (například aby nedošlo k závislosti v kruhu). V tom případě se rekční požadavek naplní, pokud vůbec, nejbližším vhodným větným členem. Ostatní větné členy, například v koordinaci, která by měla být celá závislá na předložce, se pak neopraví (a struktura zůstane chybná).

(1) *Slabé světlo vrhalo odlesky/**4** na/**-2** skleněné pulty/**1** a/**#** tmavé lesklé hlavně/**-3** zbraní*

Pravidlo se spustí na všechny předložky, které nejsou součástí složené předložky, nejsou na konci věty a není na nich závislý žádný následující větný člen. Na začátku vlastního pravidla jsou ještě odfiltrovány předložky *vzdor*, *navzdory* a *vstříc*, u nichž není závislý člen vyžadován tak přísně, a kombinace jako *hlasovat pro*, *být proti* aj.

Pravidlo postupuje od předložky doprava, dokud nenarazí na pravděpodobnou pravou vnější hranici předložkové fráze (sloveso, některé typy interpunkce, další předložka, nevhodný pád substantiva atd.). Pokud do té doby narazí na syntaktické substantivum ve vhodném pádě, ukládá si ho do zásobníku a stejně tak syntaktická adjektiva. Narazí-li před hranicí předložkové fráze na koordinaci, po níž následuje další syntaktické substantivum ve vhodném pádě, pokusí se převěsit substantiva na koordinační spojku a spojku na předložku (2) a (6). „Pokusí se“, protože nemůže vytvořit závislost v kruhu. Kdyby taková kruhová závislost hrozila, má pravidlo dvě možnosti složitější opravy struktury. Nevyjde-li ani jedna z těchto možností, není možné koordinaci na předložku zavěsit, rekční požadavky předložky uspokojí jen první vhodné substantivum.

Když od předložky po pravou hranici předložkové fráze nestojí žádná koordinační spojka nebo koordinační spojka nekoordinuje více slov se stejným pádem, jako vyžaduje předložka, ale najde se vhodné syntaktické substantivum, budou rekční požadavky předložky uspokojeny tímto syntaktickým substantivem (3) a (4). Pokud se najde pouze syntaktické adjektivum, bude na předložku převěšeno toto adjektivum a bude mu přiřazena syntaktická funkce ***ExD*** – elipsa (4). Jiné syntaktické funkce větných členů, které jsou nově závislé na předložce, se určí na základě větného členu, na němž je závislá předložka, jeho slovního druhu a dalších vlastností (valence).

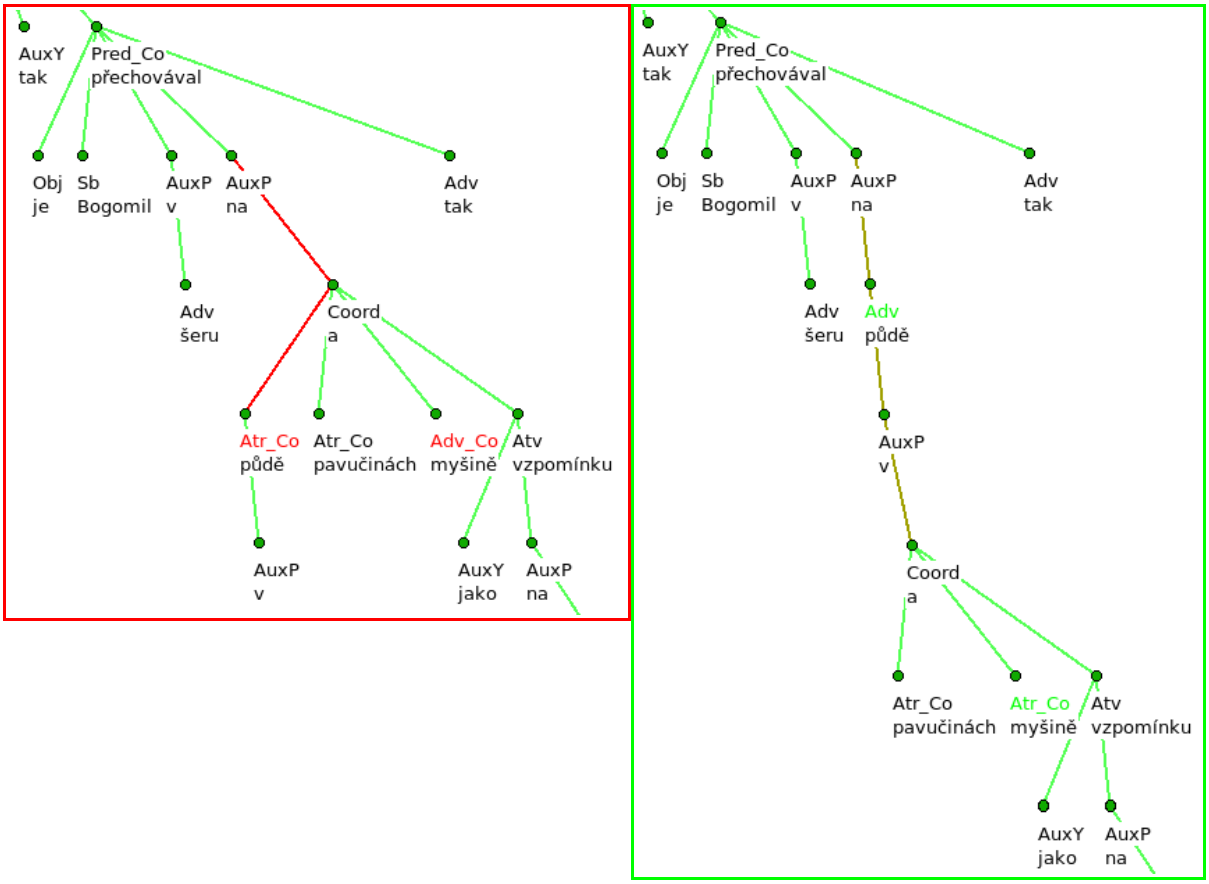
(2) *Cestu pro sebe/**Obj**/**Atr**/**1** a/**3**/**-3** čluny/**Obj**/**Atr**/**-1** si museli prosekávat .*

(3) *Nezajímala jsem se o/**-3** tebe/**Atr**/**Obj**/**3**/**-4** jako o člověka , pouze jako o matku .*

(4) *Jsou banky při jednání s/**-1** vámi/**Adv**/**Atr**/**1**/**-1** vstřícné ?*

(5) *Ty se dáš po/**-1** hlavní/**Atr**/**Exd**/**1**/**-1** a/**#** já to vezmu boční ulicí . . .*

(6) , *a tak je Bogomil přechovával v šeru na půdě v pavučinách a myšině tak jako vzpomínku na svou bývalou ženu*



Pravidlo pro doplnění neuspokojených rekčních požadavků předložek						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	181	536	49	0	0	766
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	73 %		12 %		15 %	

2.4.8 Pravidlo pro opravu závislosti v předložkové frázi se zájmenem

Ve spojeních, kde po sobě stojí předložka, zájmeno a substantivum, se parser dopouští chyb v určení závislosti, protože nedokáže rozlišit mezi zájmeny fungujícími jako syntaktická substantiva (*sebe*, *jemuž*), zájmeny fungujícími obvykle jako syntaktická adjektiva (*svému*, *nějaký*) a zájmeny, která mohou fungovat jako syntaktická substantiva i adjektiva podle kontextu (*všem*, *toho*), mimo jiné proto, že některé morfologické značky mezi těmito typy také nerozlišují (*nic* a *žádný* patří podle morfologických značek ke stejnému druhu zájmen). Parser tak i zájmena, která jsou

syntaktickými substantivy, označuje jako přívlastky následujících substantiv, která pak zavěšuje na předložku (1) a (2). K problematickému rozlišení syntaktických adjektiv a substantiv se často připojuje také chybná disambiguace pádu substantiva (3), jehož pád se chybně shoduje s pádem předložky.

(1) *Předpokládám , že pro sebe/**Atr**/**2** vhodnou postel/**Adv**/**-2** má .*

(2) *Udělal jsem pro všechny/**Atr**/**1** kotel/**Adv**/**-1** čaje s medem a rumem*

(3) *Chodil se svým otcem na ryby a vyprávěl si s/**RR--7** ním/**P5MS7**/**Atr**/**1** vtipy/**NNIP7**/**Adv**/**-2** .*

Pravidlo se zavolá, když hlavní program najde po předložce zájmeno, jež rozpozná jako syntaktické substantivum: buď je zájmeno vždy syntaktickým substantivem (*sebe*, *mě*), nebo je syntaktickým substantivem v daném kontextu v důsledku neshody s následujícím substantivem (*všechno*, *ten*). Toto zájmeno není závislé na těsně předcházející předložce, ale na následujícím substantivu.

Pravidlo změní závislost zájmene ze substantiva na předložku, substantivum pak převěsí na nejbližší sloveso. Syntaktickou funkci zájmena opraví podle pádu, předložky a případné valence slovesa. Funkci substantiva pravidlo opraví podle jeho pádu a podle valence slovesa (4), popř. podle významu substantiva, např. v (5) rozpozná substantivum jako časové určení a přiřadí funkci ***Adv***.

Pravidlo se nepokouší o opravu disambiguace, nedokáže ani spolehlivě určit, že je oprava namístě (6), někdy ale kvůli chybné disambiguaci pádu substantiva chybně určí jeho syntaktickou funkci (7): chybná oprava funkce (***Obj***) je vyznačena oranžovou barvou.

(4) *To by podle všeho/**Atr**/**Adv**/**1**/**-1** krizi/**Adv**/**Obj**/**-2**/**1** prohloubilo natolik ,*

(5) *Richard na něj/**Atr**/**Obj**/**1**/**-1** chvíli/**Adv**/**-1**/**1** hleděl nechápavě .*

(6) *vyznávají a s ním/**Atr**/**Adv**/**1**/**-1** věky/**NNIP7**/**Adv**/**1**/**-1** věků bydleti sobě vinšují*

(7) *Proto vybíhaly ženy z baráků , a ať na ně dozorkyně/**NNFP4**/**Adv**/**Obj**/**-1**/**1** křičely ,*

Pravidlo pro opravu závislosti v předložkové frázi se zájmenem						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	23	61	0	0	0	84
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	79 %		17 %		4 %	

2.4.9 Pravidlo pro opravu subjektu v předložkové frázi

Formalismus PDT umožňuje v souladu s pojetím Šmilauerovým (Šmilauer, 1966) považovat za vyjádření podmětu předložkové fráze ve výrazech přibližnosti a podílnosti: *Své lampy rozžehlo na/-**1** sta/**Sb** mušek/**Atr** v trávě* (Šmilauerův příklad citovaný v manuálu PDT). V PDT je takto označeno cca 70 výrazů (nepočítaje chyby), všechny vyjadřují přibližný počet (*okolo 500/**Sb** dělostřeleckých granátů, na třicet/**Sb** domů, přes čtrnáct/**Sb** tisíc příznivců*) nebo distribuci (*po 28 žácích/**Sb***). Pouhých 70 výskytů je ale na spolehlivé natrénování parseru málo. V trénovacích datech se navíc objevují složitější struktury jako *Na třicet/**Sb** ázerbájdžánských vojáků a dva Arménci/**Sb** byli zabiti v sobotu během bojů na severovýchodě Náhorního Karabachu*, kde je koordinována předložková fráze s nominativem a tato koordinace je závislá na slovese v plurálu.

Parser se tedy v trénovacích datech setká i s větou, která obsahující subjekt v předložkové frázi a její sloveso přitom není v singuláru. U nově anotovaných vět pak takové konstrukce považuje za přijatelné. Parser navíc není schopen ověřovat najednou větší množství podmínek, které subjekt v předložkové frázi vyžaduje.

Syntaktickou funkci ***Sb*** přiřazuje parser předložkovým frázím jen velmi zřídka (jen asi 0,2 % subjektů je podle parseru v předložkové frázi, zčásti přes chybně závislou koordinaci), ale toto přiřazení působí spíše náhodně a většinou je chybné (1) a (2). Někdy je značka přiřazena správně, chybně je ale určena závislost: substantivum vůbec nepatří do předložkové fráze (3).

(1) *Například pro/**3** francouzské banky/**Sb**/**-2** byl nárůst podvodů hlavním důvodem pro přechod*

(2) *Přes/**3** 900000/**Sb**/**-1** dolarů vynaložil otcův sponzor za jediný rok na zorganizování jeho schůzek*

(3) *Nikdy z něj jachtař/**Sb**/**-2** nebude , řekl později Arne .*

Pravidlo, jež se zaměřuje na tyto chyby, je opatrné, opravuje jen přehledné konstrukce. Hlavní program je vyvolá, když se ve větě setká s předložkovou frází, v níž je syntaktickému substantivu přiřazena funkce ***Sb***. Pravidlo ověří, zda předložková fráze jako celek splňuje podmínky pro takové označení: sloveso, na němž je předložková fráze závislá, musí být v singuláru, a jde-li o příčestí, musí být v singuláru neutra (struktura analogická výše uvedené výjimce, která se objevuje v PDT, nebyla v datech nalezena); na slovese nesmí být závislý jiný podmět, zvláště ne obvyklý podmět v nominativu; předložka řídící předložkovou frázi musí patřit mezi následující předložky: *okolo*, *kolem*, *k*, *přes*, *na* s akuzativem, *po* s lokálem; předložková fráze musí obsahovat kvantifikaci, ať už v řídícím členu fráze (*na padesát/**Sb** nejrůznějších map*/***Atr**/**-2***), nebo v přívlastku (*kolem 2000/**Atr**/**1** vzorků/**Sb**/**-2***), s výjimkou singuláru po předložce *po*, kde kvantifikace přímo vyjádřena

být nemusí.

Nesplňuje-li předložková fráze všechny tyto podmínky zároveň, změní pravidlo její syntaktickou funkci na ***Obj*** nebo ***Adv*** v závislosti na pádu substantiva, lemmatu předložky a případné valenci slovesa (4) a (5). Pravidlo nedokáže správně rozpoznat nepřesnou kvantifikaci přímého předmětu (analogie s přibližností u ***Sb***), a přiřazuje tedy funkci chybně (6). Opravu závislostí substantiv v nominativu, která jsou (správně) označena jako ***Sb***, ale chybně závisejí na předložce, by měla zajistit jiná pravidla.

(4) *Na ryby/**Sb**/**Obj** působí poškozením CNS a změnou krevního obrazu fenoly*

(5) *Likvidátor například komunikuje s centrálou výhradně přes přenosný počítač/**Sb**/**Adv** ,*

(6) *Slavia už nyní dluží firmě ENIC , která své injekce do klubové pokladny poskytuje formou půjček , kolem 400 miliónů/**Sb**/**Adv** korun .*

Pravidlo pro opravu subjektu v předložkové frázi						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	0	0	18	0	0	18
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	86 %		14 %		0 %	

2.4.10 Pravidlo pro předložkovou frázi s předložkou *o* před komparativem

Poslední dvě pravidla ověřují méně časté, specifické struktury s předložkami, u nichž parser vykazuje vyšší chybovost. První strukturou jsou předložkové fráze s předložkou *o* s akuzativem stojící před komparativem adjektiva nebo adverbia, kdy předložková fráze upřesňuje míru rozdílu ve vlastnosti vyjádřené komparativem. Parser u takovýchto obratů chybuje ve více než jedné třetině, a to jak v určení závislostí, tak v označení syntaktických funkcí (1), (2) a (3).

(1) *Nikdy nepochopím , proč Anna měla o/**-1** tolik/**Adv**/**1** hodnější/**Atr**/**1** děti než já .*

(2) *která zaznívá ve skutečnosti o/**-1** oktávu/**Atr**/**-1** níž/**Atr**/**-1** ,*

(3) *V pátek přijel Rickard a vypadal o/**-1** deset/**Adv**/**-1** let mladší/**Obj**/**-4** .*

Pravidlo se aktivuje, když v analyzované větě následuje po předložce *o* substantivum, zájmeno nebo číslovka v akuzativu a po něm komparativ. V omezené míře může mezi číslovkou v akuzativu a komparativem stát ještě genitiv plurálu substantiva. Pravidlo ověří, zda syntaktické substantivum či adverbium po předložce patří do seznamu cca 40 slov, které často upřesňují srovnání. Nejsou-li pak správně závislosti nebo syntaktické funkce (předložka závisí na komparativu, syntaktické

substantivum závisí na předložce a má funkci ***Adv***), pravidlo funkce a závislosti opraví. V případě číslovky následované substantivem v genitivu plurálu připouští pravidlo mezi substantivy a komparativy pouze omezený počet obratů s vyjádřením času (*o několik hodin později*, *o pár let starší*...), vzdálenosti (*o deset kilometrů dál*) nebo ceny (*o tisíc korun dražší*). Parser zde nechybuje v určení závislosti substantiva, ale u předložky, výjimečně i u kvantifikátoru (6). Je-li komparativ (chybně) závislý na předložce nebo na slově po předložce, určí pravidlo jako řídící uzel komparativu dosavadní řídící uzel předložky (7).

Pravidlo zachází opatrně s předložkovými frázemi závislými na slovesech s valencí *o + ak.* obsahujícími adjektiva v komparativu, v (8) tedy nezasáhne, může však v podobném kontextu chybovat u adverbií v komparativu, jako v (9). Chybně opravená struktura v příkladu (9) je však syntakticky víceznačná.

(4) *jsou hodně od sebe , s řasami o/-1/2 odstín/Atr/Adv/-1 tmavšími/-3 než vlasy .*

(5) *O/4/2 poznání/Obj/Adv/-1 dražším/1 pomocníkem jsou grily plynové ,*

(6) *Pak jsem se potkal v o/-1/3 pár/Adv/-2/-1 let/Atr/-1 starším/1 vydání .*

(7) *O/4/2 to/Obj/Adv/-1 výmluvnějšími/Obj/-2/2 se staly Neprašovy sochy .*

(8) *Nejde totiž o/-2 nic/Obj/-1 menšího než Ústavní soud , mocnou instituci ,*

(9) *když jsme ji o/6/2 to/Obj/Adv/-1 souvisleji/1 a trochu stranou požádali .*

Pravidlo pro předložkovou frázi s předložkou *o* před komparativem						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	38	9	50	0	0	97
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	94 %		4 %		2 %	

2.4.11 Pravidlo pro neprojektivní spojení předložky, adjektiva, číslovky a slovesa

Neprojektivita není pro parser nijak závažnou překážkou pro vytvoření správné závislostní struktury, v případě dílčí struktury, na niž se toto pravidlo zaměřuje, však chybuje přibližně v polovině jejích výskytů. Neprojektivní konstrukce typu *za posledních patnáct let* se v korpusu SYN2005 vyskytuje pouze asi 2600krát, pravidlo má tedy jen velmi malý efekt. V trénovacích datech je podobných struktur jen asi 20. Parser chybuje především v určení závislosti a funkce syntaktického adjektiva po předložce a před číslovkou, za jeho řídící člen považuje někdy předložku (1), jindy číslovku (2). V menší části případů je konstrukce navíc chybně disambiguovaná (3).

(1) *Pošlete na to svoje nejlepší lidi , na celých/**ExD**/**-1** čtyřiadvacet/**Adv**/**-2** hodin/**Atr**/**-1** a samozřejmě ozbrojené .*

(2) *Na potřebných/**Atr**/**1** šest/**Adv**/**-2** miliónů/**Atr**/**-1** korun se složilo ministerstvo kultury ,*

(3) *Za/**RR--2** prvních/**Atr**/**3** patnáct/**Cn--1**/**Sb**/**3** srpnových/**Atr**/**1** dnů/**Atr**/**-2** navštívilo centrum ve Falkultní nemocnici v Porubě 1714 dárců krve a plazmy .*

Hlavní program spustí toto pravidlo, když narazí na předložku s akuzativem (potenciálně s akuzativem, bez ohledu na aktuální disambiguaci), po níž následuje jedno syntaktické adjektivum v genitivu nebo lokálu plurálu (homonymní tvar), popř. více takových syntaktických adjektiv, dále číslovka *pět* a více nebo číslovka neurčitá typu *několik*, *pár* v akuzativu (popř. v nominativu, opět je to homonymní tvar), dále následují případně další syntaktická adjektiva v genitivu plurálu a konečně substantivum v genitivu plurálu.

Pravidlo ověří správnost závislostí, syntaktických funkcí a morfologických značek. Není-li struktura správně interpretována (na předložce závisí číslovka, na číslovce substantivum, na substantivu všechna syntaktická adjektiva), opraví závislosti, opraví také syntaktické funkce (4) a (5). Jsou-li ve spojení chybné morfologické značky, opraví také je (6).

(4) *Za/**6** prvních/**Atr**/**1**/**2** pět/**Adv**/**-2** měsíců/**Atr**/**-1** roku 2002 získaly firmy obchodované na londýnské burze*

(5) *v Německu se za/**6** prvních/**Atr**/**1**/**2** osm/**Cn-S4**/**Sb**/**Adv**/**4**/**-2** měsíců/**Atr**/**-1** roku 2000 zvedl knižní prodej o dvě procenta .*

(6) *Po/**RR--6**/**RR--4**/**4** celých/**AAIP6**/**AAIP2**/**Atr**/**2** šest/**Obj**/**Adv**/**5**/**-2** týdnů/**NNIP2**/**Atr**/**-1** můžeš každou hodinu věnovat akcím .*

Pravidlo pro neprojektivní spojení předložky, adjektiva, číslovky a slovesa						
počet zásahů	závislosti	závislosti a funkce	synt. funkce	funkce a morf. zn.	morf. značky	celkem
	8	3	0	0	4	15
úspěšnost oprav	úspěšná oprava		neutrální změna		negativní zásah	
	96 %		4 %		0 %	

3. Celková úspěšnost opravného programu

V tomto oddíle shrneme úspěšnost jednotlivých pravidel a počet zásahů tak, aby bylo možné odhadnout celkové zlepšení syntaktického značkování, které opravný program zajišťuje. Údaje získané z korpusu SYN2005 srovnáme také s podobnými údaji z testovacích dat PDT (e-test), kde bylo porovnání provedeno automaticky.

3.1 Úspěšnost opravného programu v korpusu SYN2005

Úspěšnost opravného programu v korpusu SYN2005 odvozujeme ze dvou doplňujících se údajů o každém aplikovaném pravidlu: počet změn v korpusu provedených každým pravidlem a odhad úspěšnosti pravidel odvozený ze vzorku aplikací pravidla.

3.1.1 Tabulka úspěšnosti jednotlivých pravidel a celého programu

V první tabulce představujeme procento úspěšnosti jednotlivých pravidel spočítané ze vzorků aplikace pravidel (100 náhodných zásahů v různých textech) a celkový počet zásahů (přepočítáno na 1 000 000 tokenů). Uvedeno je všech 26 pravidel v pořadí, v němž byla představena, spolu s odkazy na číslo oddílu, kde je popsána jejich funkce. Údaje již byly uvedeny jednotlivě u každého pravidla, zde jsou jen shrnuty do jedné tabulky.

Úspěšnost pravidel v korpusu SYN2005		Procento úspěšnosti			
	Zaměření pravidla	úspěšných oprav (%)	neutrálních změn (%)	negativních zásahů (%)	celkem
2.1.1	Koordinace nekomp. vět. členů jako #	86 %	6 %	8 %	355
2.1.2	Vedlejší věty jako #	76 %	15 %	9 %	259
2.1.3	Závislost souřadných hlavních vět	96 %	4 %	0 %	1178
2.1.4	Vztažné věty	83 %	7 %	10 %	559
2.1.5	Slovesa se spojkou *-li*	88 %	12 %	0 %	521
2.2.1	Závislosti přes hranici klauzí	91 %	7 %	2 %	565
2.2.2	Spojka *jako*	51 %	49 %	0 %	404
2.2.3	Reflexivum *se*	86 %	12 %	2 %	4423
2.2.4	Změny tagu shodných přívlastků	68 %	0 %	32 %	121
2.3.1	Shodné substantivní přívlastky	81 %	16 %	3 %	1063
2.3.2	Dva subjekty na jednom slovese	79 %	19 %	2 %	993
2.3.3	Podmět v akuzativu	68 %	24 %	8 %	887
2.3.4	Koordinace nekompat. funkcí	82 %	12 %	6 %	601
2.3.5	Předmět netranzit. slovesa	85 %	15 %	0 %	310
2.3.6	Předmět mod./fáz. slovesa	87 %	4 %	9 %	114
2.4.1	Synt. funkce v předl. frázích	70 %	29 %	1 %	7871
2.4.2	Předl. fráze závislé na slovesech	87 %	5 %	8 %	2938
2.4.3	Předl. fr. závislé na rekčních subst.	85 %	6 %	9 %	235
2.4.4	Předl. fr. závislé na nerekčních subst.	73 %	12 %	15 %	766
2.4.5	Koordinace závislé na předložkách	70 %	19 %	11 %	127
2.4.6	Víceslovné předl. výrazy	76 %	13 %	11 %	79
2.4.7	Neuspokojené předložky	76 %	9 %	15 %	319
2.4.8	Předl. fráze se zájmeny	79 %	17 %	4 %	84
2.4.9	Subjekt v předložkové frázi	86 %	14 %	0 %	18
2.4.10	Předložka *o* před komparativem	94 %	4 %	2 %	97
2.4.11	Neprojektivní konstrukce s předl.	96 %	4 %	0 %	15
	Procento úspěšnosti	79 %	17 %	4 %	100 %
	Celkem zásahů (na 1 000 000 tokenů)	19667	4293	942	24902

Opravný program tedy zasahuje přibližně 2,5 % tokenů v korpusu SYN2005, z toho úspěšných zásahů je cca 79 %, ale 4 % zásahů ovlivňují struktury negativně, a tyto negativní zásahy je nutné od úspěšných oprav odečíst. Výsledná úspěšnost je tedy cca 75 %.

3.1.2 Tabulka typu změn pro jednotlivá pravidla i celý opravný program

Ve druhé tabulce jsou představeny počty úspěšných zásahů opravného programu podle jejich typu, tj. podle toho, co bylo při zásahu opraveno. Rozlišujeme pět typů opravy: opravu, která zasahuje pouze závislostní strukturu (určení řídícího větného členu slova); opravu závislostní struktury spolu se syntaktickou funkcí slova; opravu pouze syntaktické funkce; opravu syntaktické funkce a zároveň morfologické značky; opravu morfologické značky (bez dalších změn).

Uvedená čísla jsou součinem procenta úspěšnosti (počet úspěšných zásahů snížený o počet chybných změn) a evidovaného počtu zásahů jednotlivého druhu v celém korpusu, přepočteno na 1 000 000 tokenů. Z počtu zásahů je odvozeno také předpokládané zlepšení syntaktické anotace v procentech.

Rozdělení úspěšných zásahů podle typu		Typ opravy					
	Zaměření pravidla	záv.	z.+fce	funkce	fce+t.	tagy	celkem
2.1.1	Koordinace nekomp. vět. čl. jako #	193	30	55	0	0	278
2.1.2	Vedlejší věty jako #	90	64	19	0	0	173
2.1.3	Závislost souřadných hlavních vět	1061	14	56	0	0	1131
2.1.4	Vztažné věty	151	253	4	0	0	408
2.1.5	Slovesa se spojkou *-li*	260	199	0	0	0	459
2.2.1	Závislosti přes hranici klauzí	490	13	0	0	0	503
2.2.2	Spojka *jako*	0	1	206	0	0	207
2.2.3	Reflexivum *se*	50	10	3655	0	0	3715
2.2.4	Změny tagu shodných přívlastků	0	0	0	0	44	44
2.3.1	Shodné substantivní přívlastky	532	291	6	0	0	829
2.3.2	Dva subjekty na jednom slovese	36	35	484	210	0	765
2.3.3	Podmět v akuzativu	0	8	344	0	181	533
2.3.4	Koordinace nekompat. funkcí	18	1	438	0	0	457
2.3.5	Předmět netranzit. slovesa	11	0	93	160	0	264
2.3.6	Předmět mod./fáz. slovesa	74	0	2	12	0	88
2.4.1	Synt. funkce v předl. frázích	482	9	4940	0	0	5431
2.4.2	Předl. fráze závislé na slovesech	51	4	2266	0	0	2321
2.4.3	Předl. fr. závislé na rekčních subst.	166	0	13	0	0	179
2.4.4	Předl. fr. závislé na nerek. subst.	98	0	97	0	0	195
2.4.5	Koordinace závislé na předložkách	49	2	24	0	0	75
2.4.6	Víceslovné předl. výrazy	18	31	2	0	0	51
2.4.7	Neuspokojené předložky	105	311	28	0	0	444
2.4.8	Předl. fráze se zájmeny	17	46	0	0	0	63
2.4.9	Subjekt v předložkové frázi	0	0	15	0	0	15
2.4.10	Předložka *o* před komparativem	35	8	46	0	0	89
2.4.11	Neprojektivní konstrukce s předl.	8	3	0	0	4	15
	Celkem úspěšných zásahů	3994	1332	12792	382	228	18729
	Předpokládané zlepšení v %	0,40 %	0,13 %	1,28 %	0,04 %	0,02 %	1,87 %

Z tabulky vyplývá, že většina oprav ovlivňuje pouze syntaktické funkce, chyby v určení závislosti budou opraveny jen asi v polovině procenta tokenů v korpusu (předpokládaná redukce chyb v určení závislosti z 15,9 % na 15,4 %, tj. o 3 %). Ze sond vyplývá, že dohromady opravný program redukuje chyby přibližně o 8 % (počet chyb v určení závislosti nebo syntaktické funkce se sníží z 23 % na 21,1 %).

3.2 Úspěšnost pravidel měřená na testovacích datech PDT

Úspěšnost pravidlového programu byla měřena i samostatně na testovacích datech PDT. Zde uvádíme jednak tzv. accuracy (přesnost určení řídícího uzlu a přesnost určení syntaktické funkce), jednak tabulky úspěšnosti a typů zásahů pro jednotlivá pravidla, které jsou stejné jako tabulky pro korpus SYN2005, ale výpočet se provedl automaticky, mírně odlišnou metodou.

3.2.1 „Accuracy“ použitého modelu MST parseru a opravného programu

Pro hodnocení kvality závislostního syntaktického značkování se používají dva parametry: „unlabeled accuracy“ a „labeled accuracy“. „Unlabeled accuracy“ označuje podíl správných určení řídícího uzlu u všech tokenů. „Labeled accuracy“ označuje podíl správných určení řídícího uzlu a zároveň syntaktické funkce u všech tokenů. V následující tabulce jsou uvedeny údaje pro samotný MST parser v nastavení, které jsme použili pro syntaktickou anotaci korpusu, a pro MST parser doplněný o opravný program.

e-test PDT	MST parser	MST + opravný program	rozdíl
unlabeled accuracy	84,12 %	84,48 %	0,36 %
labeled accuracy	76,95 %	78,28 %	1,33 %

Z tabulky vyplývá, že zlepšení naměřené na testovacích datech PDT je znatelně menší než předpokládané zlepšení, které jsme vyvodili ze sond v korpusu SYN2005. Možné příčiny tohoto rozdílu uvádíme dále po rozboru výkonu jednotlivých pravidel, vcelku však lze konstatovat, že zlepšení opravným programem není dostatečné, pro větší zlepšení kvality syntaktickou anotaci bude třeba hledat i jiné cesty. Podotýkáme, že samotný vstup, tj. výstup MST parseru bez opravného programu, je znatelně horší než nejlepší publikované výsledky téhož parseru. V následujících dvou tabulkách uvedeme výkon jednotlivých pravidel v e-testu PDT.

3.2.2 Tabulka úspěšnosti jednotlivých pravidel a celého programu

V následující tabulce představujeme automaticky zjištěné hodnoty úspěšnosti jednotlivých opravných pravidel na e-testu PDT. Úspěšnost se vypočítává automaticky porovnáním s manuálně anotovanými daty, v některých ohledech je přísnější než v předchozí tabulce (nevhodná oprava syntakticky zcela nesmyslné struktury se může zařadit i mezi negativní zásahy aj.). Počet zásahů jsme přepočítali na 1 milión tokenů, aby byl srovnatelný s výše uvedenými tabulkami.

Úspěšnost pravidel v e-testu PDT		Procento úspěšnosti			
	Zaměření pravidla	úspěšných oprav (%)	neutrálních změn (%)	negativních zásahů (%)	celkem
2.1.1	Koordinace nekomp. vět. členů jako #	44 %	40 %	16 %	464
2.1.2	Vedlejší věty jako #	50 %	21 %	29 %	302
2.1.3	Závislost souřadných hlavních vět	88 %	8 %	4 %	799
2.1.4	Vztažné věty	75 %	15 %	10 %	994
2.1.5	Slovesa se spojkou *-li*	73 %	19 %	8 %	281
2.2.1	Závislosti přes hranici klauzí	77 %	13 %	10 %	518
2.2.2	Spojka *jako*	43 %	57 %	0 %	400
2.2.3	Reflexivum *se*	67 %	33 %	0 %	1760
2.2.4	Změny tagu shodných přívlastků	55 %	45 %	0 %	119
2.3.1	Shodné substantivní přívlastky	55 %	28 %	17 %	1631
2.3.2	Dva subjekty na jednom slovese	59 %	41 %	0 %	1199
2.3.3	Podmět v akuzativu	61 %	37 %	2 %	983
2.3.4	Koordinace nekompat. funkcí	46 %	50 %	4 %	583
2.3.5	Předmět netranzit. slovesa	80 %	20 %	0 %	108
2.3.6	Předmět mod./fáz. slovesa	60 %	40 %	0 %	54
2.4.1	Synt. funkce v předl. frázích	56 %	41 %	3 %	8867
2.4.2	Předl. fráze závislé na slovesech	79 %	21 %	0 %	2408
2.4.3	Předl. fr. závislé na rekčních subst.	83 %	10 %	7 %	324
2.4.4	Předl. fr. závislé na nerekčních subst.	53 %	27 %	20 %	324
2.4.5	Koordinace závislé na předložkách	76 %	24 %	0 %	184
2.4.6	Víceslovné předl. výrazy	40 %	0 %	60 %	54
2.4.7	Neuspokojené předložky	65 %	16 %	19 %	810
2.4.8	Předl. fráze se zájmeny	100 %	0 %	0 %	22
2.4.9	Subjekt v předložkové frázi	100 %	0 %	0 %	11
2.4.10	Předložka *o* před komparativem	92 %	8 %	0 %	130
2.4.11	Neprojektivní konstrukce s předl.	100 %	0 %	0 %	22
	Procento úspěšnosti	62 %	32 %	6 %	100 %
	Celkem zásahů (na 1 000 000 tokenů)	14558	7527	1266	23351

Jak je z tabulky vidět, poměr úspěšných a chybných zásahů jednotlivých pravidel testovaných na e-testu PDT se značně liší od výsledků manuálního testování vzorků z korpusu SYN2005. Pravidlo pro víceslovné předložkové výrazy svými zásahy v konečném důsledku syntaktickou anotaci zhoršuje (na základě průzkumu trénovacích dat lze však soudit, že důvodem zhoršení může být nejednotné značkování víceslovných předložek v testovacích datech, které zvyšuje počet rozdílů automaticky anotovaného textu oproti manuálnímu značkování testovacího vzorku). Jiná pravidla

ve většině svých zásahů mění syntaktickou anotaci, aniž by se jim podařilo anotaci zlepšit.

3.2.3 Tabulka typu změn pro jednotlivá pravidla i celý opravný program

Ve druhé tabulce opět rozdělujeme úspěšné zásahy podle typu. Počet úspěšných zásahů se počítá jako rozdíl úspěšných zásahů a negativních, chybných zásahů. Je-li počet chybných zásahů v dané kategorii vyšší než počet zásahů správných, je číslo záporné (**červeně** zvýrazněno). Počet zásahů je přepočítán na 1 milión tokenů stejně jako v předchozí tabulce.

Rozdělení úspěšných zásahů podle typu		Typ opravy					
	Zaměření pravidla	záv.	z.+fce	funkce	fce+t.	tagy	celkem
2.1.1	Koordinace nekomp. vět. čl. jako #	129	0	0	0	0	129
2.1.2	Vedlejší věty jako #	**-10**	53	22	0	0	65
2.1.3	Závislost souřadných hlavních vět	648	0	22	0	0	670
2.1.4	Vztažné věty	159	476	12	0	0	647
2.1.5	Slovesa se spojkou *-li*	54	129	0	0	0	183
2.2.1	Závislosti přes hranici klauzí	346	0	0	0	0	346
2.2.2	Spojka *jako*	0	0	173	0	0	173
2.2.3	Reflexivum *se*	46	0	1142	0	0	1188
2.2.4	Změny tagu shodných přívlastků	0	0	0	0	65	65
2.3.1	Shodné substantivní přívlastky	430	172	13	0	0	615
2.3.2	Dva subjekty na jednom slovese	43	55	425	179	0	702
2.3.3	Podmět v akuzativu	0	22	366	0	184	572
2.3.4	Koordinace nekompat. funkcí	**-16**	11	253	0	0	248
2.3.5	Předmět netranzit. slovesa	0	0	22	64	0	86
2.3.6	Předmět mod./fáz. slovesa	32	0	0	0	0	32
2.4.1	Synt. funkce v předl. frázích	13	13	4628	0	0	4654
2.4.2	Předl. fráze závislé na slovesech	23	11	1845	0	0	1879
2.4.3	Předl. fr. závislé na rekčních subst.	216	0	32	0	0	248
2.4.4	Předl. fr. závislé na nerek. subst.	22	0	86	0	0	108
2.4.5	Koordinace závislé na předložkách	87	0	54	0	0	141
2.4.6	Víceslovné předl. výrazy	0	**-10**	0	0	0	**-10**
2.4.7	Neuspokojené předložky	107	236	35	0	0	378
2.4.8	Předl. fráze se zájmeny	0	22	0	0	0	22
2.4.9	Subjekt v předložkové frázi	0	0	11	0	0	11
2.4.10	Předložka *o* před komparativem	22	0	97	0	0	119
2.4.11	Neprojektivní konstrukce s předl.	22	0	0	0	0	22
	Celkem úspěšných zásahů	2373	1190	9238	243	249	13293
	Zaznamenané zlepšení v %	0,24 %	0,12 %	0,92 %	0,02 %	0,02 %	1,33 %

Z tabulky je zřejmé, že počet úspěšných zásahů v e-testu PDT je výrazně nižší než v korpusu SYN2005 (cca o 40 %). Na e-testu se ukazuje celkové zlepšení úspěšnosti o 1,33 %, na korpusu SYN2005 jsme vypočítali zlepšení úspěšnosti o 1,87 %. Rozdíly lze vysvětlovat mnoha způsoby, jež nelze ověřit, protože e-test musí zůstat pro manuální kontrolu nepřístupný. Některé z možných důvodů horších výsledků opravného programu na e-testu PDT uvedeme v následujícím odstavci.

E-test PDT je řádově menšího rozsahu než korpus SYN2005, na němž jsme opravný program vyvíjeli a testovali (e-test je cca tisíckrát menší). Je možné, že některé jevy, které opravný program úspěšně koriguje v korpusu SYN2005, se v e-testu neobjevují. Při testování opravného programu na trénovacích datech PDT jsme narazili i na občasné chyby manuální syntaktické anotace, nelze tedy vyloučit ani chyby v e-testu. Zásahy opravného programu, které nevhodným způsobem řešily zcela nesmyslnou syntaktickou strukturu, se na e-testu započítaly jako negativní, přestože výsledná struktura je „méně chybná“. Pokud by například pravidlo, které vyhledává a odstraňuje dva nekoordinované subjekty závislé na jednom slovese, špatně zvolilo a opravilo nevhodný subjekt, vznikla by sice lepší struktura s jedním ***Sb*** a jedním ***Obj*** či ***Pnom***, ale při testování by se projevila jen změna ze správného ***Sb*** na chybný ***Obj*** či ***Pnom***.

Skutečné výsledky by podle těchto hypotéz mohly být mírně lepší, než jak byly automaticky zjištěny, tyto úvahy však nelze ověřit a do výsledků započítat (testovací data musí zůstat „slepá“). Za prokázané považujeme tedy to, že opravný program zlepšil výsledky „labeled accuracy“ MST parseru o 1,33 %.

3.3 Závěr

Zlepšení o 1,33 % v případě automaticky syntakticky anotovaného korpusu není zanedbatelné, ale zaostává za našimi očekáváními. Vstupní data navíc obsahovala o něco vyšší procento chyb než nejlepší publikované výsledky MST parseru (Novák et al. 2007), ty se nám ale dosud nepodařilo reprodukovat, celkovou úspěšností se tak naše výsledky řadí jen mezi lepší průměr syntaktické anotace češtiny, nejlepší dosažené výsledky jsme nepřekonali.

Přesto nelze vývoj tohoto opravného programu považovat za slepou uličku vývoje: jak ukážeme dále, opravný program lze dále vyvíjet a lze ho také aplikovat na výstupy parserů s lepšími výsledky. Přínos opravného programu u lépe nastavených parserů bude sice nižší, ale celkový výsledek by měl přesáhnout i nejlepší dosažené výsledky automatické syntaktické anotace.

4. Další možný rozvoj opravného programu

Výsledky opravného programu zaostávají za původním očekáváním, přesto znatelně zlepšují

výsledky stochastického parseru. Jednou z možných cest k dalšímu zlepšení kvality syntaktické anotace je tedy další rozvoj opravného programu, který by měl zahrnout opravu současných pravidel, rozšíření seznamů slov, jichž pravidla využívají, doplnění nových pravidel a zapojení frazémového programu.

4.1 Oprava chybujících pravidel a doplnění algoritmů

Při testování jednotlivých pravidel na nových datech bylo zjištěno, že některá pravidla se poměrně často dopouštějí chyb: zasahují do správných struktur (zřídka) nebo chybně opravují špatné struktury a dále tak zhoršují jejich stav. Chybovost pravidel (procento negativních zásahů) se pohybuje mezi 0 % a 32 % (na e-testu PDT až 60 %). V pravidlech, která často chybují, je nutné upravit algoritmus tak, aby k chybám docházelo méně nebo vůbec. K dispozici je dostatečně rozsáhlý korpus, na kterém lze zásahy testovat, do pravidel lze snadno přidat omezení nebo složitější rozhodovací algoritmy. Podrobnější rozpracování pravidel se vyplatí pouze u těch, jež se aplikují dostatečně často.

Některá pravidla jsou naopak příliš opatrná a opravují jen jasně přehledné struktury, i když dokážou identifikovat mnohem více chybných struktur: nemají-li bezpečný algoritmus pro opravu, nezasahují do struktury vůbec. Pokud je ale takováto struktura (identifikovaná jako chybná, ale neopravovaná) dostatečně častá, vyplatí se upravit algoritmus tak, aby pravidlo opravovalo i další podtypy chyby, aniž by přitom dělalo více chyb.

4.2 Doplnění dalších pravidel

Rozbor výsledků parseru a také analýza výsledků opravného programu ukázala, že oblastí, na něž by se pravidla mohla zaměřit (často se opakující typ chyby s jasným opravným algoritmem), je i po implementaci prvních 26 pravidel ještě dost, zvláště v oblasti závislostí okolo koordinací a interpunkce. Lze však předpokládat, že přínos rostoucího množství pravidel bude klesat, případné zdvojnásobení počtu pravidel by pravděpodobně vedlo k menšímu než dvojnásobnému zlepšení parsingu. Zvažujeme tedy zapojit jen několik dalších pravidel, u nichž předpokládáme dostatečně vysokou návratnost a malou chybovost a jejichž oblast působnosti již byla zmapována. U některých pravidel jsou již připravené opravné algoritmy, stačí pravidlo jen implementovat, jiná pravidla jsou nutná proto, že chybné struktury blokují opravy, které by mohla provést pravidla již implementovaná.

4.3 Rozšíření používaných seznamů

Seznamy, jež pravidla využívají pro práci s valencí sloves a adjektiv, pro určení reflexivity, pro zpracování skupin typu *pan Novák* aj., pocházejí zčásti ze starších zdrojů, jako je projekt pravidlové

morfologické disambiguace, zčásti byly nově vyvozeny z korpusových dat (korpusy SYN2005 a SYN2010). Při takovém poloautomatickém zpracování byly do seznamů zařazeny pouze dostatečně frekventované výskyty. Rozšíření seznamů o méně frekventovaná slova umožní spolehlivě opravovat větší procento chybných struktur.

V případě valence bude navíc nutné zjemnit valenční seznam. Tento seznam dosud obsahuje pouze informaci, že určité sloveso má valenci s jistým pádem nebo pádem a předložkou (*toužit po 6*), popř. u slovesa, kde je časté spojení s okolnostním určením, obsahuje seznam informaci, že spojení pádu či předložky a pádu s daným slovesem je adverbiálního typu (*schovat za 4*). Pro spolehlivější určení závislostí předložkových frází by však bylo třeba rozlišovat valenci alespoň ve dvou stupních (fakultativní a obligatorní): jestliže se například určité sloveso s obligatorní předložkovou valencí objeví ve stejné klauzi s odpovídající předložkovou frází, téměř vždy bude předložková fráze záviset na slovesu (*toužit po 6*, *usilovat o 4*), zatímco vazba jiného slovesa s fakultativní předložkovou frází může být mnohem slabší (*začít s 7*, *bránit v 6*), i když obě předložkové fráze mohou být označeny jako ***Obj***, tj. jako valenční předmět slovesa. Díky jemnějšímu rozlišení valenčních požadavků sloves by méně chybovala například pravidla zaměřená na změnu závislosti předložkových frází (v případě volby mezi závislostí na slovesu a na substantivu aj.).

5. Závěr

Výsledky opravného programu zaostávají za původním očekáváním, přesto znatelně zlepšují výsledky stochastického parseru. Ke znatelnému zvýšení spolehlivosti automatické syntaktické anotace bude nutná jak další práce na rozvoji tohoto programu, tak lepší nastavení samotných parserů.