V. Kůrková et al. (Eds.): ITAT 2014 with selected papers from Znalosti 2014, CEUR Workshop Proceedings Vol. 1214, pp. 61–67
http://ceur-ws.org/Vol-1214, Series ISSN 1613-0073, © 2014 M. Lopatková, J. Mírovský, V. Kuboň

Gramatické závislosti vs. koordinace z pohledu redukční analýzy

Markéta Lopatková, Jiří Mírovský a Vladislav Kuboň

Univerzita Karlova v Praze, Matematicko.fyzikální fakulta, Ústav formální a aplikované lingvistiky
Malostranské nám. 25, Praha 1, 118 00, Česká republika
{vk,lopatkova,mirovsky}@ufal.mff.cuni.cz

Abstrakt: Tento příspěvek se věnuje identifikaci zajímavých konstrukcí v syntakticky anotovaném korpusu (Pražském závislostním korpusu, PDT) metodou automatické redukční analýzy. Rozšiřujeme zkoumané konstrukce zejména o koordinační a apoziční vztahy, které mají zřetelně ne-závislostní charakter a vedou tedy k zobecnění používané redukční metody. Přinášíme klasifikaci zkoumaných konstrukcí a soustřeďujeme se na popis a analýzu jednotlivých jazykových jevů, které při zpracování působí problémy.

Tato studie je motivací pro formální modelování metod zpracování přirozeného jazyka.

1 Motivace

Jedním ze dvou základních a obecně uznávaných způsobů reprezentace syntaktických vztahů ve větách přirozeného jazyka je závislostní strom. Tento způsob reprezentace má v české lingvistice dlouhou tradici (na rozdíl od druhého rozšířeného typu stromu – složkového). Přestože vztah závislosti (neboli vztah mezi řídícím a závislým větným členem, jako je např. přísudek a jeho předmět či podstatné jméno a jeho přívlastek) je velmi důležitý, doplňují jej i dva další základní syntaktické vztahy, jmenovitě slovosled (tj. lineární pořadí slov ve větě) a ‚zmnožení' dvou nebo více větných členů.[1]

Zatímco v předchozích studiích jsme se soustřeďovali na vztahy závislosti a slovosledu [2], v tomto článku zkoumáme i vztahy koordinační a apoziční, věnujeme se tedy případům, kdy je jedna syntaktická pozice ‚zmnožena'. Ukázali jsme již, že závislostní relace lze s úspěchem definovat pomocí redukční analýzy [3, 4], což jsme ověřili na větách z Pražského závislostního korpusu (PDT) [5]. Na ně jsme aplikovali poloautomatickou proceduru doplněnou o následnou ruční kontrolu. Této metody se držíme i v tomto článku. Umožňuje nám verifikovat náš teoretický koncept prostřednictvím reálných dat.

Smyslem našeho experimentu je jednak získat hlubší vhled do syntaxe přirozeného jazyka, jednak prostřednictvím automatizovaného postupu identifikovat určitá nestandardní či problematická místa syntaktické anotace dat, se kterými pracujeme. V otázce syntaktických vlastností jazyka náš přístup umožňuje oddělit jednotlivé jevy a zkoumat je jak jednotlivě, tak i ve vzájemné interakci. V oblasti anotace vycházíme z jednoduchého předpokladu (potvrzeného předchozími experimenty), že pokud určitá konstrukce nejde zpracovat automatickými pravidly redukční analýzy, signalizuje to možnou nekonzistentní nebo nevhodně zvolenou anotaci (například dva odlišné jevy anotované shodnými značkami – v takovém případě obvykle automatická redukční analýza nedokáže oba případy rozlišit).

V kontextu závislostní lingvistiky byly vztahy mezi závislostí a slovosledem studovány zvláště v Melčukově teorii Smysl↔Text: jeho přístup zaměřující se na určování závislostních vztahů a jejich formální popis je shrnut zejména v práci [6]. Alternativní formální popis závislostní syntaxe můžeme najít v práci [7]. Náš přístup k danému problému naproti tomu vychází z české lingvistické tradice reprezentované zejména v knize [8].

Protože základním nástrojem, který používáme ke studiu výše zmíněných jevů, je *redukční analýza*, musíme ji na tomto místě alespoň stručně představit: zhruba řečeno, pokud jedno ze slov, která tvoří potenciální dvojici řídícího a závislého slova, může být z věty odstraněno, aniž by se změnily distribuční vlastnosti celého páru (tj. jeho schopnost objevovat se ve stejném syntaktickém kontextu), potom je toto slovo považováno za závislé (modifikující druhý členu páru). Takto můžeme postupovat u tzv. endocentrických konstrukcí, kde lze jedno slovo redukovat, aniž by se změnil možný syntaktický kontext (např. *malý stůl –> stůl*; *Jdi domů! –> Jdi!*). Pro exocentrické konstrukce, kde žádné slovo vypustit nelze (jako *Petr potkal Marii.*, kde *Petr potkal.* má jiné vlastnosti) lze použít analogický princip na úrovni slovních druhů [8].[2]

Důvod pro využívání redukční analýzy je jednoduchý – umožňuje rozdělit proces syntaktické analýzy věty do jednotlivých, dobře definovaných a dobře oddělených kroků, a tím zároveň dovoluje zkoumat jednotlivé jevy odděleně. Metoda redukční analýzy byla podrobně popsána v článcích [3, 4], její formální model založený na restartovacích automatech je představen v článcích [9, 10, 11]. Jedním z cílů tohoto článku je také poskytnout materiál a motivaci pro další formální modelování metod zpracování přirozeného jazyka, viz např. [12].

[1] Tesnière [1] jednak rozlišuje mezi lineárním a strukturním pořadím, jednak dělí strukturní vztahy na ty, které dnes označujeme jako závislostní (‚connexion'), a na vztahy koordinační (‚junction').

[2] Zhruba řečeno, protože existují bezpředmětná slovesa, budeme objekt (předmět) považovat za závislý na slovese; protože existují bezpodmětná slovesa, považujeme subjekt (podmět) za závislý na slovese.

2 Redukční analýza

V této kapitole popíšeme základní myšlenku naší metody používané pro analýzu vět. *Redukční analýza (RA)* je založena na postupném zjednodušování analyzované věty. Definuje možné posloupnosti redukcí věty – každý krok RA spočívá v *odstranění* alespoň jednoho slova ze vstupní věty (operace 'delete'); ve specifických případech je odstraňování doprovázeno *přemístěním* slova na jinou slovoslednou pozici (operace 'shift').

Uveďme nyní základní omezení, která uplatňujeme na redukční analýzu:

(i) přirozené omezení nutící zachovávat jednotlivé slovní tvary, jejich morfologické charakteristiky a jejich povrchové závislostní vztahy;

(ii) omezení vyžadující zachování správnosti (gramaticky správná věta musí zůstat správná i po provedeném zjednodušení);

(iii) aplikace operace přesunutí je omezená na případy, kdy je vynucena principem zachování správnosti RA (ii).

Povšimněme si, že pořadí redukcí odpovídá závislostním vztahům mezi jednotlivými větnými členy, a tedy jejich závislostní stromové reprezentaci tak, jak je to popsáno v pracích [4, 11]. Zhruba řečeno, při redukcích postupně vypouštíme slova, která jasou reprezentována listy (případně podstromy) závislostního stromu.

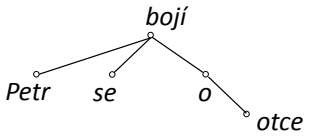
Pokusme se ukázat základní principy RA na příkladu české věty (1).

Příklad:
(1) *Petr se bojí o otce.*

Ve větě (1) lze subjekt *Petr* považovat za závislý (viz princip analogie), proto ho při RA lze odstranit (v závislostním stromě je reprezentován listem, viz obrázek 1). Ovšem tento krok by vedl k porušení správnosti věty, musí tedy být doprovázen změnou slovosledu *Petr se bojí o otce.* $\rightarrow_{delete}$ * *se bojí o otce.* $\rightarrow_{shift}$ *bojí se o otce.*. Dále lze vypustit spojení *o otce*, které je ve stromě znázorněno podstromem (výběr řídícího uzlu je dán technickými pravidly, viz [3]). Klitika *se* je podle principu analogie považována za závislý člen (neboť existují slovesa bez klitiky, např. *odpovědět*).

3 Data z Pražského závislostního korpusu

Ačkoliv základní princip redukční analýzy je jednoduchý a rodilým mluvčím daného jazyka obvykle nečiní potíže větu zredukovat až k jejímu řídícímu slovesu, automatické modelování tohoto postupu je poměrně obtížné. Tento rozdíl je způsoben tím, že člověk při redukční analýze může využít (a využívá) toho, že větě rozumí, umí oddělit méně



Obrázek 1: Závislostní strom věty (1).

podstatné větné členy a postupně dojít až k úplné redukci. Pokusíme-li se o automatickou RA, musíme porozumění nějakým způsobem nahradit. Jednou možností je využití syntakticky anotovaných dat, která v sobě určitým způsobem obsahují porozumění, vložené do anotace člověkem-anotátorem.

V našich experimentech využíváme data z Pražského závislostního korpusu 3.0 (PDT, viz [5]).[3] Syntaktická struktura jednotlivých vět z korpusu – zachycená závislostními stromy (vždy právě jeden strom pro každou větu) – poskytuje základní informace nutné pro úspěšné provedení automatické RA.

PDT obsahuje velmi podrobnou anotaci téměř 49 500 českých vět (v experimentech využíváma pouze data s anotací na všech rovinách PDT). Anotace je provedena na více úrovních, z nichž je pro naše účely nejdůležitější úroveň analytická. Ta popisuje (povrchovou) syntaktickou strukturu pomocí tzv. analytických funkcí. V našich experimentech pracujeme pouze s trénovacími daty (43 955 vět) a zbylé věty (tzv. 'etest') ponecháváme stranou jako testovací množinu pro budoucí evaluaci.

Pro reprezentaci větné stavby se v PDT používá kořenový strom. Vztahy mezi členem řídícím a závislým, vytvářejícími větnou dvojici, jsou znázorněny jako vztahy mezi dvěma uzly stromu, kde uzel reprezentující řídící člen je rodičem a uzel reprezentující závislý člen je potomkem. Jejich spojnice (hrana ve stromě) odpovídá (prototypicky) syntaktickému vztahu závislosti mezi rodičovským uzlem a jeho potomkem.

V předchozích experimentech, popsaných v článcích [2, 13], jsme se záměrně vyhýbali větám, které obsahovaly koordinace a apozice, protože tyto jevy by byly pro počáteční fázi výzkumu příliš komplikované. Podstatou koordinace a apozice je ‚zmnožování' jednotlivých větných členů nebo jednotlivých klauzí. Při koordinaci jde o zřetězení více entit (*otec a matka*), popř. dějů (*přišel, viděl a zvítězil*), při apozici jde o několikeré označování jedné a téže entity (*Karel, král český*), viz [14].

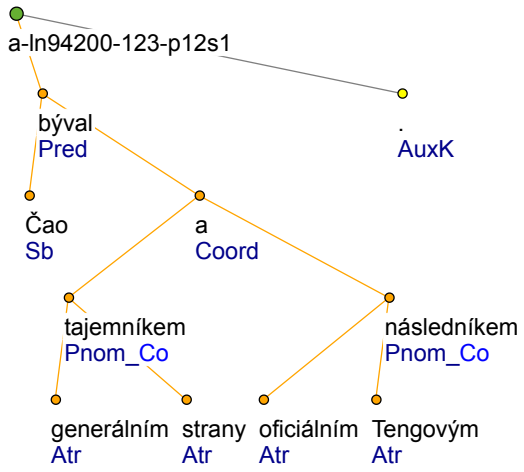
Protože koordinace ani apozice nepředstavují jevy, které by v sobě obsahovaly přirozenou závislost, bývají v syntaktických stromech reprezentovány různým způsobem [15]. V závislostních stromech používaných v PDT jsou tyto jevy zachyceny tzv. spojovací konstrukcí (viz příklad (2) a strom na obrázku 2). Kořenem souřadných struktur je umělý ‚spojovací' uzel; z důvodů čistě technických je jeho lexikální hodnotou lemma souřadicí spojky (či lemma výrazu signalizujícího apozici). Vlastní koordinované/apo-

[3] http://ufal.mff.cuni.cz/pdt3.0/

nované výrazy jsou spojeny s tímto kořenem hranami, které mají ne-závislostní charakter. Celá spojovací konstrukce je potom spojena s řídícím uzlem celé koordinační struktury další (nezávislostní) hranou.

Příklad:
(2) *Čao býval generálním tajemníkem strany a oficiálním Tengovým následníkem.*

Závislostní strom věty (2), tak jak je zachycen na analytické rovině PDT, je uveden na obrázku 2. Například hrana mezi uzly *býval* a *Čao* reprezentuje závislostní vztah predikátu a subjektu. Spojovacím výrazem pro koordinaci je zde uzel pro spojku *a*, koordinovanými členy výrazy *tajemníkem* a *následníkem* a řídícím uzlem celé konstrukce spona *býval*.



Obrázek 2: Závislostní strom věty (2) podle pravidel PDT (dále již pro jednoduchost neuvádíme technický kořen stromu (obsahující ID dané věty) a koncovou interpunkci).

4 Automatická redukční analýza na datech PDT

Návrh automatické redukční analýzy vyžaduje pečlivý návrh redukčních pravidel tak, aby byl v každém kroku analýzy zajištěn požadavek na zachování správnosti redukovaných vět. Postupuje se ‚zdola nahoru' – postupným redukováním listů závislostního stromu z PDT (který nahrazuje porozumění dané větě), přičemž nejprve se vždy redukují uzly bezprostředně sousedící se svým řídícím uzlem; následuje redukce uzlů spojených s řídícím uzlem projektivní hranou.

Přitom je zachovávána důležitá podmínka na zachování neprojektivity – nelze redukovat uzel tak, aby z věty ‚zmizela' neprojektivní konstrukce.

4.1 Pravidla pro RA bez koordinace a apozice

V předchozí etapě projektu jsme představili soubor pravidel pro automatickou RA [13, 2] – ukázalo se, že pro věty bez koordinační či apoziční struktury automatická RA dobře koresponduje s lingvistickou analýzou zachycenou v PDT. Intuitivní RA musela být zjemněna tak, aby byly správně zpracovány zejména následující jevy, které přesahují čistě závislostní vztahy – typicky tam, kde je nutné pro zachování správnosti redukované věty brát v úvahu slovosled:

klitiky: klitiky (zejm. *se/si*, *by*, krátké tvary zájmen (*mu*, *ji* apod.)) vyžadují určité postavení ve větě, typicky za první přízvučnou pozicí; toto pravidlo je nutno při RA zohlednit;

srovnání: srovnávací konstrukce (typicky uvozené výrazy *jako*, *než*, dále *coby*, *jakoby*, *jakožto*) vyžadují zvláštní zpracování, neboť jde často o eliptické konstrukce (*Petr má větší auto než Pavel.* = *Petr má větší auto [než je auto, které má] Pavel.*) a jako takové se vyznačují složitou analýzou (která je závislá na podkladové lingvistické teorii);

neprojektivity: při automatické analýze jsme se omezili na zpracování projektivních konstrukcí, neboť neprojektivity odhalují interakci slovosledu a závislostních vztahů, která je dále studována, viz zejm. [13];

předložky, pomocná slovesa apod.: v RA se redukují v jednom kroku vždy celé větné členy (např. předložka+podstatné jméno v redukci *Přišli do školy.* –> *Přišli.*); protože v PDT každému slovu, včetně těchto ‚pomocných' slov, odpovídá jeden uzel, musí se redukovat několik uzlů stromu najednou;

slova překračující závislostní vztahy: RA byla obohacena o pravidla pro zpracování zdůrazňujících slov (např. *zejména*, *také*, *i* apod.) a o technická pravidla pro zpracování interpunkce, grafických symbolů (závorky, uvozovky atd.) apod.

4.2 Pravidla pro RA se zpracováním koordinace a apozice

Další krok při zobecňování RA spočívá v zaměření se na paradigmaticky odlišné konstrukce, a to konstrukce koordinační (a apoziční), které jsou charakterizovány zmnožením příslušné syntaktické pozice.

Pravidla pro automatickou RA obohacenou o zpracování koordinace a apozice:

0. Ve vstupní větě jsou zpracovány jevy nevstupující do (dané) koordinační či apoziční struktury (včetně společného rozvití koordinovaných větných členů), které lze zpracovat podle pravidel shrnutých výše.
Tento krok se provádí jako mezikrok po každém úspěšném zpracování koordinační či apoziční struktury.

1. Při zpracování koordinačních a apozičních struktur se vždy redukuje spojovací výraz (typicky koordinační spojka či interpunkce) spolu s koordinovanými (či aponovanými) výrazy.

2. Všechny koordinované (či aponované) členy se redukují v jediném kroku analýzy. Teoreticky nelze omezit počet koordinovaných členů; přestože v datech PDT se obvykle koordinuje 2-5 členů, v jednom případě se koordinuje 57 členů (televizní program syntakticky strukturovaný do věty), u apozic jde až o 15 členů.

3. Zdůrazňovací syntaktické částice a všechna pomocná slova, interpunkce, grafické symboly apod. se redukují zároveň v jednom kroku RA (podle dříve stanovených pravidel).

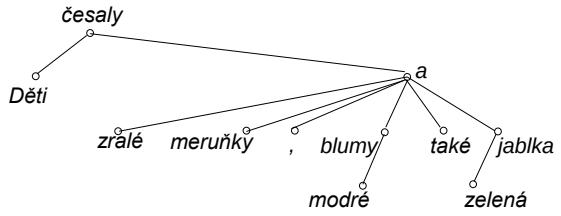
4. Koordinace a apozice dovolují zanořování. Opět jde teoreticky o neomezený počet zanoření, viz [16]; v datech PDT se vyskytlo 6 úrovní zanoření koordinací a apozic (přehled sportovních výsledků).

I v této fázi se soustřeďujeme pouze na automatické zpracování projektivních konstrukcí; neprojektivní konstrukce jsou analyzovány ručně.

Kroky 1-3 redukční analýzy ilustrujeme na následujícím příkladu (3).

Příklad:
(3) *Děti česaly zralé meruňky, modré blumy a také zelená jablka.*

(Kde přídavné jméno *zralé* je analyzováno jako společné rozvití slov *meruňky*, *blumy* a *jablka*, viz obr. 3.)



Obrázek 3: Závislostní strom věty (3) (podle pravidel PDT, bez technického kořene a koncové interpunkce).

Krok 0: Redukují se všechna slova, která nevstupují do koordinace (získáme tím ‚kostru koordinace' se společným rozvitím:
–> *česaly zralé meruňky, blumy a také jablka.*
Dále se redukuje společné rozvití koordinace:
–> *česaly meruňky, blumy a také jablka.*

Krok 1 a 2:
- redukuji se všechny členy koordinace *meruňky*, *blumy*, *jablka*
- redukuje se spojovací uzel, tedy koordinační spojka *a*

Krok 3: Zároveň se v témže kroku RA zpracovávají zdůrazňovací slova a interpunkce:
- redukuje se zdůrazňovací *také*
- redukuje se interpunkční čárka
- redukuje se koncová interpunkční tečka
–> *česaly*

Manuální analýza automatických redukcí získaných z PDT vedla k dalším zpřesněním pravidel, která se týkala zejména následujících jevů.

Další pravidla pro automatickou RA obohacenou o zpracování koordinace a apozice:

5. Zpracovávají se věty s koordinační spojkou, kde se však ‚koordinuje' jediná klauze (či větný člen) a kde tedy koordinační spojka plní funkci odkazu k předcházejícímu kontextu. Např.
Nemáme proto potíže ***se získáváním trhu pro své výrobní odpady***.
–> *Nemáme proto potíže*
(Dále se neredukuje kvůli neprojektivitě.)

6. Víceslovné spojky a syntaktické částice vztahující se k celé koordinované klauzi (typicky v PDT zachycované jako potomek spojovacího výrazu) se redukují v jednom kroku se spojovacím výrazem. Např.:
Jsou ***buď nedostupná, nebo nedostačující***.
–> *Jsou*
(Redukce probíhá v jediném kroku: spojovacím uzlem je podle pravidel PDT koordinační spojka *nebo*, druhá část spojovacího výrazu *buď* musí být redukována zároveň (pravidlo 6); koordinované členy *nedostupná* a *nedostačující* se redukují podle kroku 1, interpunkční čárka podle kroku 3.)

Další pravidla se již se netýkají specificky koordinací:

7. Adekvátněji se redukují konstrukce s modálním slovesem (např. *měly tvořit*) a s verbonominálním predikátem (např. *Je učitelem*, viz též větu (2)): vnitřní struktura těchto konstrukcí je zjednodušována až na závěr RA.

8. Emocionální a rytmizující částice *mi*, *vám*, *si*, *to*, *ono* apod. jsou redukovány kdykoli v průběhu redukce, i pokud jde o klitiky.

5 Lingvistická analýza zajímavých příkladů

Podívejme se nyní podrobněji na jevy, u kterých buď (povrchová) redukční analýza nedává uspokojivé výsledky, nebo odhaluje zajímavé syntaktické konstrukce.

Pozice klitiky v koordinované klauzi. Podle Encyklopedického slovníku češtiny [17] je příklonka (klitika) slovo (zpravidla krátké), které nemá vlastní přízvuk, tvoří přízvučný celek (takt) se slovem předcházejícím. Klitiky mají v češtině pevné slovosledné postavení (zpravidla) za prvním přízvučným celkem (první pozicí) ve větě (tzv. Wackernagelova pozice). Syntaktický popis první pozice je poměrně komplexní [18], pro naše účely stačí konstatování, že klitika nesmí stát na prvním místě ve větě.

Poněkud nejasná situace nastává v případě klitiky ve druhé (a další) koordinované klauzi. Souřadicí spojky stojí podle českých gramatik mimo koordinované klauze, proto by po nich měl následovat přízvučný celek a teprve potom pozice pro klitiky. To platí např. pro některé spojky slučovací (*a*, *i*) a odporovací (*ale*):
Cesta sice něco stojí, ale zákazníci se o kvalitě produkce přesvědčí na vlastní oči.
–> *cesta sice stojí, ale* ***zákazníci*** *se přesvědčí*

*–> * cesta sice stojí, ale **se** přesvědčí*

*Diskuse by se přenesla [...], a **zcela** by vybočila z hranic ekonomie.*
*–> * Diskuse by se přenesla [...], a **by** vybočila z hranic ekonomie.*

Oproti tomu u řady dalších souřadicích spojek (např. slučovací/vylučovací *nebo*, příčinné *neboť*, důsledkové *(a) proto*, vysvětlovací *vždyť*) může tato spojka sama tvořit první přízvučný celek, klitika tedy může (ale nemusí) následovat bezprostředně po ní:
*Film můžeme považovat za plně autorský, neboť **Chabrol** si k němu napsal i scénář [...].*
–> můžeme, neboť si k němu napsal i scénář

Podrobnější lingvistický rozbor slovosledných omezení ve vztahu ke klitikám lze nalézt v knihách [19, 20]. Nejsme si však vědomi, že by problematika koordinační spojky a postavení klitiky byla pro češtinu v lingvistické literatuře popsána.

Koordinace s elidovanými větnými členy. Podobně jako u srovnávacích konstrukcí též u koordinačních spojení často dochází k elipse části syntaktické struktury (tzv. aktuální elipsa). V (povrchových) závislostních stromech PDT se elipsa nerekonstruuje, větné členy syntakticky závislé na členu vypuštěném se zavěšují tam, kde by visel člen vypuštěný (a označují se speciální analytickou funkcí ExD).

Takové případy ovšem způsobují při automatické RA problémy, neboť při redukci může docházet k porušení plynulosti (či alespoň stylistické souvislosti) redukovaných vět (i když z hlediska čistě formální syntaxe jsou správně):
*[...], Flintstoneovi sice nepřinášejí zábavu **náročnou**, ale ani nevkusně prostoduchou. (= prostoduchou zábavu)*
–> ??[...], Flintstoneovi sice nepřinášejí zábavu, ale ani nevkusně prostoduchou.

*[...] jsou a zdaleka ne **tak** časté jako ředění zmrzlin. (= jako je časté ředění zmrzlin)*
–> ?? [...] jsou a zdaleka ne časté jako ředění zmrzlin.

Možným řešením je zcela vyloučit eliptické konstrukce související s koordinací z automatického zpracování (vzhledem k takto vzniklým nepravidelnostem obvykle nelze věty tohoto typu úspěšně automaticky redukovat až na základní predikativní strukturu).

Konstrukce s nejasnou závislostí – typ *do konce roku*. U jistých konstrukcí, především u některých časových určení (např. *těsně před Vánocemi*, *185 minut týdně*) a místních určení (*dva kilometry od řeky*), bývá někdy těžké rozhodnout, zda jsou na sobě nezávislá, či zda jedno rozvíjí druhé.

PDT v těchto případech vychází z poněkud technických pravidel anotačního manuálu [21]. Metoda RA zde nedává spolehlivé výsledky, viz např.
*V Bratislavě by mělo vzniknout **do konce roku**, stejně jako v New Yorku a Bruselu.*
–>?? V Bratislavě by mělo vzniknout do konce, stejně jako v New Yorku a Bruselu.
–>?? V Bratislavě by mělo vzniknout do roku, stejně jako v New Yorku a Bruselu.

*Není povinen se **každý měsíc** hlásit ve zprostředkovatelně práce, a proto je vyřazen ze statistiky.*
*–> * Není povinen se měsíc hlásit ve zprostředkovatelně práce, a proto je vyřazen ze statistiky.*
*–> * Není povinen se každý hlásit ve zprostředkovatelně práce, a proto je vyřazen ze statistiky.*

Vzhledem k poněkud nejasné syntaktické struktuře těchto jevů při vyhodnocování automatické RA od případných takto vzniklých nekorektností odhlížíme.

Víceslovné výrazy a ‚pojmenované entity‘. V datech PDT nejsou v (povrchových) stromech zachyceny tzv. pojmenované entity, což jsou např. názvy osob (*Petr Novák*, *Čao C'jang*), zvířat (*Alík*), lokací (Hradec Králové), institucí (*Česká národní banka*, *Koh-i-Noor*) apod., a víceslovné výrazy (např. *státní zástupce*, *šelma kočkovitá*) – jejich anotace podléhá stejným pravidlům jako anotace obecných jmen. (Adekvátní anotace je zachycena až na tektogramatické rovině.)

Automatická RA tedy v těchto případech nezíská z dat dostatečné informace a oba tyto typy výrazů jsou zpracovávány jako obecné konstrukce, což s sebou nese poněkud problematické redukce výrazů, které by měly být zpracovávány v jednom kroku, např.:
*Kupříkladu sedmdesátiletý **Čchiao Š'** či pětasedmdesátiletý **Čao C'jang**.*
*–> Kupříkladu sedmdesátiletý **Š'** či pětasedmdesátiletý **Čao C'jang**.*
[. . .]
*–> **Š'** či pětasedmdesátiletý **Čao C'jang**.*
*–> **Š'** či pětasedmdesátiletý **Čao jang**.*
*–> **Š'** či pětasedmdesátiletý **jang**.*

Vzhledem k tomu, že tyto jevy se primárně netýkají syntaxe, ale dostupné slovníkové informace, při návrhu automatické RA od chyb ve zpracování pojmenovaných entit a víceslovných výrazů odhlížíme.

Redukce valenčních doplnění. Při návrhu (povrchové) redukční analýzy vycházíme z povrchově syntaktické struktury vět (uchovávané na tzv. analytické rovině PDT), která nepracuje s informací o valenční charakteristice jednotlivých plnovýznamových slov a o tzv. vypustitelnosti jejich valenčních doplnění (tato informace je doplňována až na tektogramatické rovině, která je už ovšem od povrchové syntaxe poměrně vzdálena). Proto ani RA nemůže pracovat s valenční informací.

Tato skutečnost má zásadní dopad na (povrchovou) automatickou RA: protože RA nemá k dispozici valenční slovník, redukuje postupně všechna objektová a adverbiální doplnění sloves a atributivní doplnění substantiv, bez

ohledu na jejich (ne)valenční charakter a vypustitelnost. Ačkoli typicky jsou závislé členy vypustitelné, v řadě případů dochází k zásadnímu posunutí významu či k porušení plynulosti (či dokonce správnosti) redukovaných vět. Následující příklady ilustrují posun významu (*jít o něco* –> *jít*) a neúplnou strukturu (redukce adjektivního objektu v postpozici).

*Jde **o činorodost** a právě ona [...] dělá ze všech protagonistů Genu [...] elitu.*
–> Jde a právě ona [...] dělá ze všech protagonistů Genu [...] elitu.

*[...] odrážejí tendenci vývoje oproštěného **od vlivů i od hodnot**.*
*–> * [...] odrážejí tendenci vývoje oproštěného.*

Další problém spojený s valenční charakteristikou jednotlivých slov spočívá v pořadí vypouštění během RA. Pravidla pro vypouštění např. slovesných doplnění nejsou v českých gramatikách (alespoň podle nám dostupných informací) formálně zpracována, přesto je zřejmé, že zde platí nějaká omezení; např. vypouštění dativních doplnění či doplnění realizovaných předložkovou skupinou by typicky mělo předcházet vypouštění akuzativního doplnění (a to i v případě aktuální elipsy), viz následující příklady:
*[...] poskytují **služby a činnost buď přímo nebo prostřednictvím specializovaných institucí.***
–> ?? [...] poskytují buď přímo nebo prostřednictvím specializovaných institucí.
–> [...] poskytují služby a činnost.

*Občané **si** mohou **vyšetření** objednat v hygienických stanicích.*
–> ?? Občané si mohou objednat v hygienických stanicích.
–> Občané mohou vyšetření objednat v hygienických stanicích.

Vzhledem k nevyjasněným pravidlům a omezením na pořadí vypouštění jednotlivých doplnění[4] při automatické RA od stylistických pochybení či nekorektností způsobených nesprávným pořadím vypouštění valenčních doplnění odhlížíme.

6 Závěrečné shrnutí

Hlavním smyslem našich experimentů bylo ověřit použitelnost automatické redukční analýzy na složitější jazykové jevy (koordinace, apozice) a vytipovat určité problematické konstrukce, které v budoucnu poslouží jako materiál jak pro další lingvistický výzkum, tak i pro další zjemnění či modifikaci samotné metody redukční analýzy a pro formální popis vlastností přirozených jazyků. Experimenty ukázaly, že i koordinace a apozice se ve většině případů dají redukovat automaticky. Zároveň se podařilo objevit několik problematických konstrukcí, z nichž zejména spojení koordinace a postavení klitik či pořadí vypouštění slovesných doplnění představují problémy, které zatím nebyly uspokojivě lingvisticky zpracovány.

[4] Podle [22] hrají roli nejen povrchové formy doplnění, ale i typ valenčního vztahu a obligatornost doplnění, tedy informace náležející do valenčního slovníku, a tedy na tektogramatickou rovinu (nikoli na rovinu povrchové syntaxe).

Grantová podpora

Práce na tomto tématu je podpořena z grantu GAČR číslo P202/10/1333. Tento článek využívá jazyková data vyvinutá a/nebo distribuovaná v rámci projektu MŠMT ČR LINDAT/CLARIN (projekt LM2010013).

Reference

[1] Tesnière, L.: Eléments de syntaxe structurale. Librairie C. Klincksieck, Paris (1959)

[2] Kuboň, V., Lopatková, M., Mírovský, J.: A Case Study of a Free Word Order. In: Proceedings of PACLIC 2013. (2013)

[3] Lopatková, M., Plátek, M., Kuboň, V.: Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In: Proceedings of TSD 2005. Volume 3658 of LNAI., Berlin Heidelberg, Springer-Verlag (2005) 140–147

[4] Lopatková, M., Plátek, M., Sgall, P.: Towards a Formal Model for Functional Generative Description: Analysis by Reduction and Restarting Automata. The Prague Bulletin of Mathematical Linguistics **87** (2007) 7–26

[5] Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M.: Prague Dependency Treebank 2.0. LDC, Philadelphia, PA, USA (2006)

[6] Mel'čuk, I.A.: Dependency in language. In: Proceedings of DepLing 2011, Barcelona (2011) 1–16

[7] Gerdes, K., Kahane, S.: Defining dependencies (and constituents). In: Proceedings of DepLing 2011, Barcelona (2011) 17–27

[8] Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Reidel, Dordrecht (1986)

[9] Jančar, P., Mráz, F., Plátek, M., Vogel, J.: On monotonic automata with a restart operation. Journal of Automata, Languages and Combinatorics **4** (1999) 287–311

[10] Otto, F.: Restarting Automata. In: Recent Advances in Formal Languages and Applications, Studies in Computational Intelligence. Volume 25., Berlin, Springer-Verlag (2006) 269–303

[11] Plátek, M., Mráz, F., Lopatková, M.: (In)Dependencies in Functional Generative Description by Restarting Automata. In: Proceedings of NCMA 2010. Volume 263 of books@ocg.at., Wien, Austria, Österreichische Computer Gesellschaft (2010) 155–170

[12] Martin Plátek and Dana Pardubská and Markéta Lopatková: On Minimalism of Analysis by Reduction by Restarting Automata. In Morrill, G., Muskens, R., Osswald, R., Richter, F., eds.: Formal Grammar 2014. Volume 8612 of LNCS., Berlin Heidelberg, Springer-Verlag (2014) 155–170

[13] Kuboň, V., Lopatková, M., Mírovský, J.: Automatic Processing of Linguistic Data as a Feedback for Linguistic Theory. In Castro, F., Gelbukh, A., González, M., eds.:

Proceedings of the 12th Mexican International Conference on Artificial Intelligence (MICAI 2013). Volume 8265 of LNCS., Berlin Heidelberg, Springer-Verlag (2013) 252–264 volume 1.

[14] Šmilauer, V.: Novočeská skladba. SPN, Praha (1966)

[15] Štěpánek, J.: Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat). PhD thesis, MFF UK, Prague (2006)

[16] Oliva, K.: Linguistics behind the Mirror. In Lopatková, M., ed.: Information Technologies – Applications and Theory, Košice, Slovakia, Univerzita Pavla Jozefa Šafárika v Košiciach (2011) 1–6

[17] Karlík, P., Nekula, M., Pleskalová, J., eds.: Encyklopedický slovník češtiny. Nakladatelství Lidové noviny, Praha (2002)

[18] Hana, J.: Czech Clitics in Higher Order Grammar. PhD thesis, The Ohio State University (2007)

[19] Běličová, H., Uhlířová, L.: Slovanská věta. Euroslavica, Praha (1996)

[20] Běličová, H., Sedláček, J.: Slovanské souvětí. Academia, Praha (1990)

[21] Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A., Štěpánek, J., Pajas, P., Kárník, J.: Anotace na analytické rovině. návod pro anotátory. Technical Report TR-2004-23, UFAL MFF UK (2004)

[22] Lopatková, M.: O homonymii předložkových skupin. (Co umí počítač?). Karolinum, Praha (2003)