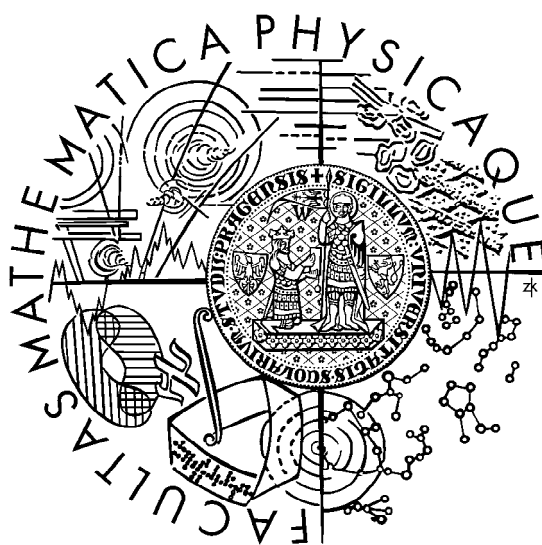


Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Nguy Giang Linh

Návrh souboru pravidel pro analýzu anafor v českém jazyce

Ústav formální a aplikované lingvistiky
Vedoucí diplomové práce: *Doc. RNDr. Jan Hajič, Dr.*
Studijní program: *Informatika*

Ráda bych poděkovala vedoucímu své diplomové práce Doc. RNDr. Janu Hajičovi, Dr. za jeho cenné připomínky a rady. Také bych mu chtěla poděkovat za poskytnuté téma, jež mi pomohlo porozumět hlouběji jazyku, který bych si přála, aby se jednou stál mým druhým mateřským jazykem.

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

V Praze dne 7.11.2006

Nguy Giang Linh

Obsah

Obsah	i
Úvod	v
Kapitola 1 Lingvistické základy k anafoře	1
1.1 První seznámení s anaforou	1
1.2 Další rozlišení anafor podle anaforických výrazů	3
1.2.1 Jmenno-frázová anafora	4
1.2.2 Zájmenná anafora	4
1.2.3 Nulová anafora	6
1.2.4 Přísluvečná anafora	7
1.3 Komplikované druhy anafor	8
Kapitola 2 Analýza anafor	10
2.1 Předpoklady k určení anafor	10
2.2 Známé způsoby analýzy anafor	12
2.2.1 Hobbsův algoritmus se stromovým vyhledáváním (1978)	12
2.2.2 BFP algoritmus založené na teorii centrování (1987)	13
2.2.3 Lappin & Leassův algoritmus se systémem vah (1994)	13
2.2.4 Mitkovův robustní algoritmus, který nepotřebuje mnoho znalosti světa	14
2.3 Strojové učení jako nástroj pro analýzu anafor	14
2.3.1 Řešení od Aone a Bennetta	15
2.3.2 Řešení od McCarthyho a Lehnerta	15
2.3.3 Řešení od Soon, Ng a Lima	15
2.4 Další způsoby analýzy anafor	16
Kapitola 3 Anafora v Pražském závislostním korpusu	17
3.1 Pražský závislostní korpus	17
3.1.1 Morfologická rovina	18

3.1.2	Analytická rovina	18
3.1.3	Tektogramatická rovina	18
3.1.4	Data v PDT	21
3.2	Koreference v PDT	22
Kapitola 4	Analyzátor české anafory	24
4.1	Osobní zájmenná anafora	24
4.1.1	Řešení pomocí strojového učení	24
4.1.2	Řešení pomocí ručně psaných pravidel	26
4.1.3	Výsledky obou řešení pro osobní zájmennou anaforu	30
4.2	Nulová anafora kontroly	32
4.2.1	Podmínky shody funktora kandidáta a funktora anaforu s kategorií prarodiče	35
4.2.2	Výsledek AČA pro nulovou anaforu kontroly	42
4.3	Reflexivní zájmenná anafora	42
4.3.1	Výsledek pro reflexivní zájmennou anaforu	44
4.4	Relativní zájmenná anafora	44
4.4.1	Řešení pro relativa kromě <i>což</i>	44
4.4.2	Řešení pro <i>což</i>	46
4.4.3	Výsledek pro relativní zájmennou anaforu	47
4.5	Demonstrativní zájmenná anafora	47
4.5.1	Výsledek pro demonstrativní zájmennou anaforu	49
4.6	Nulová anafora reciprocit	49
4.6.1	Výsledek pro nulovou anaforu reciprocit	52
4.7	Doplňek	52
4.7.1	Doplňek vyjádřený substantivem, adjektivem, zájmenem nebo číslovkou	52
4.7.2	Doplňek vyjádřený slovesem nebo lematem <i>rád, sám</i>	54
4.7.3	Výsledek pro doplňky	57
Kapitola 5	Závěr	58
Literatura		60
Příloha A	Příklady anafory v dalších jazycích	62

Příloha B	Seznam kontroly a reciprocity	63
Příloha C	Ukázka C4.5	67
Příloha D	Obsah CD-ROM	69

Název práce: *Návrh souboru pravidel pro analýzu anafor v českém jazyce*

Autor: *Nguy Giang Linh*

Katedra (Ústav): *Ústav formální a aplikované lingvistiky*

Vedoucí diplomové práce: *Doc. RNDr. Jan Hajič, Dr.*

e-mail vedoucího: *hajic@ufal.ms.mff.cuni.cz*

Abstrakt: *S rostoucí důležitostí počítačového zpracování přirozeného jazyka narůstá i množství výzkumů na téma automatické analýzy anafory. Příspěvkem k výzkumu této problematiky je rovněž naše diplomová práce, jejímž cílem je vytvořit soubor pravidel pro analýzu anafory v českém jazyce. Vytvořený soubor pravidel obsahuje jak ručně psaná pravidla, tak i pravidla vznikající pomocí systému strojového učení C4.5. K trénování a testování pravidel byla použita anotovaná data z Pražského závislostního korpusu, ve kterém je zachycena zájmenná anafora, kontrola, reciprocita a závislostní vztah doplňků. Právě těmto druhům anafory je věnována naše práce. Vyhodnocení pravidel je provedeno standardními metodami pro hodnocení úplnosti a přesnosti.*

Klíčová slova: *anafora, koreference, PDT, C4.5*

Title: *Proposal of a set of rules for anaphora resolution in Czech*

Author: *Nguy Giang Linh*

Department: *Institute of Formal and Applied Linguistics*

Supervisor: *Doc. RNDr. Jan Hajič, Dr.*

Supervisor's e-mail address: *hajic@ufal.ms.mff.cuni.cz*

Abstract: *With the increasing importance of natural language processing there is growing number of research with the theme automatic anaphora resolution.. The contribution to the research on this problem is also this thesis. The aim of the work is to propose a set of rules for anaphora resolution in Czech. The created set of rules consists of handwritten rules as well as rules developed with the aid of machine learning system C4.5. For the rules training and testing were used anoted data from the Prague Dependency Treebank, in which following types of anaphora are captured: pronominal anaphora, control, reciprocity and dependency relation of adjuncts. Our work is focused on these types of anaphora. The evaluation of the rules is done with standard methods for interpretation of recall and precision.*

Keywords: *anaphora, coreference, PDT, C4.5*

Úvod

Anafora je chápána jako zpětné odkazování k předchozím výrazům v textu. Analýza anafory se pak zabývá hledáním těchto předcházejících výrazů. Zmíněná analýza anafory hraje důležitou roli v aplikacích zpracování přirozeného jazyka např. automatický překlad nebo textová sumarizace.

Cílem této práce je vytvoření souboru pravidel pro analýzu anafory v českém jazyce, jehož správnost a úspěšnost budou ověřeny pomocí existujících dat z Pražského závislostního korpusu 2.0. Pražský závislostní korpus je sbírka lingvisticky anotovaných dat a dokumentů, na kterých je aplikován Funkční generativní popis, vyvíjený pražskými lingvisty od konce šedesátých let. Anotační schéma PDT 2.0 se skládá ze tří rovin: morfologické, analytické a tektogramtické. V rovině tektogramatické je zachycena zájmenná anafora, kontrola, reciprocita a závislostní vztah doplňků, a proto se naše práce bude věnovat právě těmto druhům anafory.

Práce je organizována následovně:

- první kapitola nabízí lingvistické základy k anafoře a její druhy
- druhá kapitola seznamuje s předpoklady k analýze anafory a známé způsoby jejího řešení
- třetí kapitola popisuje Pražský závislostní korpus a existující práce na analýze anafory v Pražském závislostním korpusu
- čtvrtá kapitola se zabývá naší vlastní implementací analýzy anafory nazvanou AČA (analyzátor české anafory) a jeho vyhodnocováním
- práci zakončuje závěr, který shrnuje výsledky práce a naznačuje její další možný vývoj

Kapitola 1

Lingvistické základy k anafoře

1.1 První seznámení s anaforou

- (1.1) a. Otec – Tak už je tady...
b. Matka – Kdo?
c. Otec – Synáček.
d. Matka (vzdech)
e. Otec – Máš tu synáčka...
f. Matka – Snad je taky tvůj, ne?
g. Otec – Vyloučeno. Můj syn by takhle pozdě domů nechodil.
h. Matka – Můj tedy také ne.

(Ivan Kraus: Má rodina a jiná zemětřesení)

Text (diskurz) je nespecifikovaný jazykový projev, tvořený z uspořádaných, souvislých vět. Vztahy v něm obstarávající jeho souvislost a srozumitelnost zajišťují **koherence** a **koheze**. Koherencí textu se rozumí celková soudržnost a spojitost v hloubkové rovině textu, a kohezi pak konkrétní realizace koherence pomocí lexikálních, syntaktických a gramatických prostředků v povrchové struktuře textu.

V přirozeném jazyce se často setkáváme s jevy působícími v rovině diskurzu. Vezměme v úvahu diskurz v následujícím příkladu:

- (1.2) *Otec* vždycky tvrdil, že opery nesnáší. Říkal, že *mu* na opeře vadí hlavně ten zpěv.

Slova *otec* a *mu* označuje jednu osobu, která je otcem hlavního hrdiny v knize. Obě slova se zúčastňují procesu, tzv. **reference**, kdy mluvčí používá různé výrazy, aby odkázal k osobám, předmětům nebo situacím reálného světa. Výrazu v přirozeném jazyce, který hraje úlohu odkazu, se podle Jurafského a Martina ([Jurafsky et Martin, 2000]) nazývá

odkazující výraz, a subjekt, na který je odkazováno, je **odkazovaný**. Tedy *otec* a *mu* jsou odkazující výrazy a pan otec v reálném světě je jejich odkazovaný. Vztah mezi dvěma odkazujícími výrazy je nazýván **koreference**, která je jedním z prostředků koheze. Jsou-li odkazující výrazy k jednomu samému odkazovanému dva nebo více, tvoří společně **koreferenční řetěz**.

Reference se obecně dělí na referenci **endoforickou** a **exoforickou**. Endoforická reference (endofora) je odkaz jednoho výrazu k jinému výrazu uvnitř téhož textu. Kdežto exoforická reference (deixe, exofora) je odkaz ke skutečností mimotextovým.

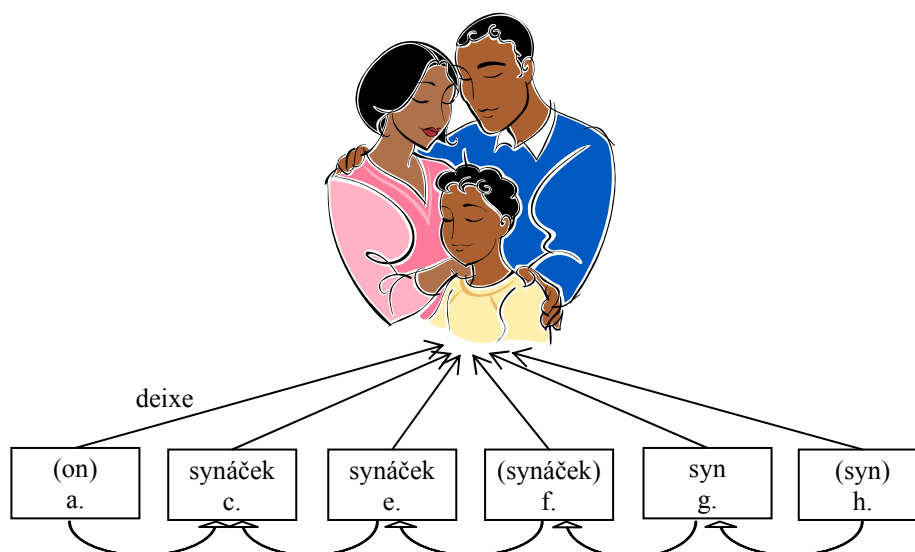
Endofora se dále ještě dělí na **anaforu** a **kataforu**. Anafora (z řeckého αναφορά, kde ανα – zpět, nahoru; φορα – nosit, vést) je odkaz k dříve zmíněnému výrazu v diskurzu, kterému se říká **antecedent**. Anaforickému výrazu se nazývá **anafor** (v Jurafském a Martinovi [Jurafsky et Martin, 2000] také anaphoric). V závislosti na umístění anaforu a jeho antecedentu ve větách jsou rozlišeny **intravětná**, **intervětná** a **extravětná** anafora. Intravětná anafora nastává, když antecedent a jeho anafora se nacházejí ve stejné klauzi, intervětná anafora se uplatňuje v posloupnosti dvou či více klauzí tvořících souvětí a extravětná v rámci textu ([Palek, 1988]). V zahraniční literatuře se setkáváme jen s dvěma typy větných anafor, a to s intravětnou anaforu pro větu a intervětnou pro text.

Opačným a řídkým případem anafory je **katafora**, referující v textu dopředu, k tomu, co teprve bude zmíněno. Referovanému výrazu se v tomto případě říká **postcedent**.

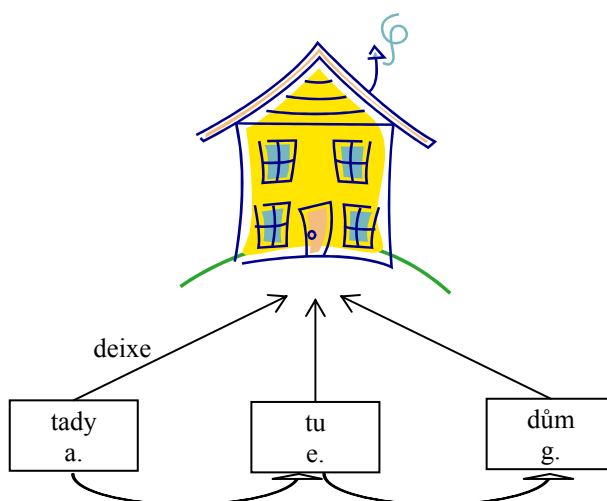
V Kučové a kol. ([Kučová et al., 2003]) se používají ještě dva termíny: **koreferující** pro anafor nebo katafor, a **koreferovaný** pro antecedent nebo postcedent.

V české lingvistice je také známé členění koreference na **gramatickou** a **textovou**. Gramatická koreference je ta, u které je zpravidla na základě gramatických pravidel možné určit antecedent. U textové koreference není odkazování dáno gramaticky, ale na základě kontextu ([Kučová et al., 2003]). Gramatickou koreferencí může být intravětná nebo intervětná anafora. Textová koreference se objevuje v rámci celého textu.

Nejčastějším jazykovým prostředkem vyjadřujícím antecedent je nezájmenná jmenná fráze. Společně s kterýmkoli druhem anaforických výrazů tvoří **nominální** anaforu, jednu z nejdůležitějších tříd anafory, která byla ve světě zpracování přirozeného jazyka rozsáhle zkoumána.



Obr. 1.1: Příklad na deixi, kataforu, anaforu a koreferenční řetěz (rozbor úvodního ilustračního diskurzu)



Obr. 1.2: Příklad na deixi, kataforu a koreferenční řetěz (rozbor úvodního ilustračního diskurzu)

1.2 Další rozlišení anafor podle anaforických výrazů

V této části kapitoly budou představeny základní druhy anaforických výrazů. A protože přirozený jazyk je nesmírně bohatý a každý jazyk má nějaké své zvláštní rozmanité prostředky, nebude tento výčet konečný.

1.2.1 Jmenno-frázová anafora

V jmenno-frázové anafoře figuruje jmenná fráze jako odkaz. Jmenné fráze odkazují k antecedentům, se kterými se shodují plně nebo částečně nebo jsou jejich synonymy. Jmenné fráze mohou být také dalšími popisy odkazovaného.

- (1.1) Juventus Turín měl před sebou ještě dva ligové zápasy, ale *Pavel Nedvěd* se už připravoval na mistrovství Evropy. "Když už bylo jasné, že nemůžeme skončit líp než třetí, začal jsem dělat na kondici," poznamenal *kapitán národního mužstva*.

Pozor na následující případ, kdy se nejedná o anaforu:

- (1.2) Šedí papoušci byli šokováni, že jsou tam také žlutí papoušci.

1.2.2 Zájmenná anafora

Osobní zájmenná anafora

Osobní zájmenná anafora obsahuje všechna zájmena v 3. osobě (kromě pleonastického *ono*). Zájmena v 1. a 2. osobě jsou častěji používána v deixi a jako anafory v přímé řeči, ale nejsou zahrnuta v analýze anafory.

Reflexivní zájmenná anafora

Anafory bývají reflexivní zájmena *se*, *si* a *svůj*. Vzhledem ke své vlastnosti jsou zájmena *se*, *si* často katafory. O anaforu se nejedná v případech, kdy reflexiva tvoří součást ustáleného spojení jako je „sám od sebe/svým způsobem/být svůj/několikrát za sebou/...“ (více v [Kučová et al., 2003]).

- (1.1) *Dramaturgie si* z půlstoleté Stahuljakovy tvorby jednostranně vybrala skladby z dvacátých let.
- (1.2) *Sázková kancelář Fortuna* přepustila *svůj* menšinový podíl, který měla ve slovenské kanceláři Terno, výhradně občanům Slovenska.

- (1.3) V těchto dnech *si* totiž *chorvatská kulturní veřejnost* spolu se znalci jihoslovanské hudby z ostatní Evropy připomíná dvacáté výročí skladatelova úmrtí.

Reflexivní zájmena většinou odkazují k subjektu hlavní klauze, kromě takových význačných případů zájmena *svůj*, kdy *svůj* odkazuje k nevyjádřenému podmětu zapuštěné klauze:

- (1.4) *Profesor_i* prověřil *asistent_j* přednést *svůj_{i,j}* referát na konferenci. ([Panevová, 1991]))

Posesivní zájmenná anafora

Posesivní zájmenná anafora používá zájmena *jeho*, *její* a *jejich* k odkazování.

- (1.5) *Sennovy* problémy postřehl v závodě i Michael Schumacher, pozdější vítěz VC: "Jezdil jsem za *Ayrtonem* a v šestém kole jsem si všiml, že *jeho* vůz v té zatáčce maličko ztrácel stabilitu, o kolo později se tam vyboural."

Ukazovací zájmenná anafora

Možnými anafory jsou všechna ukazovací zájmena kromě pleonatického *to*, které je popisováno dále.

- (1.6) Především společenským bombónkem se stal písňový cyklus op. 4 v podání *ambasadorovy choti*, *sopranistky Ivanky Stahuljak*. *Ta* zaujala spíše výrazovou stránkou projevu a niterností uměleckého prožitku než kvalitou či technikou *svého* hlasu.

Anaforické *to* často odkazuje k většímu úseku textu:

- (1.7) *Rozprava o podobě reformy veřejných financí bude zahájena ve středu. Všechna jednání proběhnou za zavřenými dveřmi.* Lidovým novinám *to* sdělil včera ministr financí.

Pleonastické *to/ono*

Pleonastické *to/ono* nikam neodkazuje, vyskytuje se v ustálených slovních spojeních (\emptyset je užitečné/ \emptyset je důležité/vypadá *to*, že.../je *tomu*/...), je intenzifikátorem (*to* ale *prší/ono* je jedno/...) nebo je obsahově vyprázdňené (*to* máte těžké). (více v [Kučová et al., 2003])

Vztažná zájmenná anafora

Vztažné anafory jsou vyjádřeny zájmeny (který/kdo/co/...) a zájmennými příslovcí (kdy/kde/jak/...), které uvádějí vedlejší větu vztažnou a většinou se vážou k antecedentu v řídicí větě. V případě spojovacího výrazu *což* může být antecedentem část věty, celá věta nebo i posloupnost vět.

(1.8) Členové družstva ČR a SR se měli sejít v kompletním složení *včera, kdy* z turnajů v Gstaadu a Ósace přicestovali Karel Nováček s Petrem Kordou.

(1.9) *Eva Urbanová se v oněch pěti číslech pokoušela prosadit především dramatický výraz, čímž si ve vysoko posazených áriích - zejména v jejich závěrečných pasážích - poněkud "zablokovala" svou techniku.*

1.2.3 Nulová anafora

Nulová anafora (elipsa) nastává, když je anafor ve větě vypuštěný (bude označen v příkladech ‘ \emptyset ’). Ten, který není v textu vyjádřen žádným slovem ani frází, ale přesto je ze smyslu výpovědi pochopen. Mitkov tvrdí: „Vzhledem k tomu že jednou z vlastností a výhod anafory je schopnost redukovat množství informace, která je prezentována zkratkovými lingvistickými formami, může elipsa být nejs sofistikovnější variantou anafory.“²

Nulová zájmenná anafora

Nejrozšířenějším druhem české anafory je nulová zájmenná anafora. Stejně jako u zájmenné anafory neobsahuje zájmena v 1. a 2. osobě.

¹ \emptyset je symbol zastupující nulový anafor (viz kapitola 1.2.3 Nulová anafora)

² Mitkov, R.: *Anaphora Resolution*, Longman, Londýn 2002, str. 12

- (1.1) *Můj strýček Melik* byl patrně nejhorší farmář pod sluncem. \emptyset Měl v duši příliš mnoho představivosti a poezie, než aby *mu* farmaření vynášelo.

Nulová anafora kontroly

Nulová anafora kontroly je koreferenční vztah mezi jistým členem valenčního rámce řídicího slovesa a nevyjádřeným valenčním členem závislého infinitivního slovesa. Platí i pro případy, kdy se řídicí sloveso nebo závislé infinitivní sloveso nominalizuje.

- (1.2) Novelu zákona o malé privatizaci včera *sněmovně* doporučil \emptyset schválit rozpočtový výbor.
- (1.3) Ale berní *úřady_i* na sebe daly dlouho čekat a nemyslím, že by za současného obsazení \emptyset_i měly u *plátců_j* daní velkou váhu a \emptyset_j respekt \emptyset_i .

Nulová anafora recipicity

Reciprocita je sémantické-syntaktický vztah vzájemnosti. Nulový anafor recipicity odkazuje k aktantům řídicího slovesa.

- (1.4) *Jan a Marie* se setkaly \emptyset .

Nulová jmenná anafora

- (1.5) Mám rád detektivní *knihy*, ale dostal jsem jednu \emptyset vědecko-fantastickou.
- (1.6) Jedna *dívka* přišla včas, ta druhá \emptyset o pár minut později.
- (1.7) Na louce se pásly *ovce*. Některé \emptyset byly černé, některé \emptyset bílé.

Nulová slovesná anafora

- (1.8) *Vyhraj Golf GTi* nebo \emptyset týden na Floridě nebo \emptyset v Paříži.³

1.2.4 Přísllovečná anafora

- (1.1) V roce 1994 jsem přijela do *České republiky*. Moc se mi *tady* líbí.

³ Mitkov, R.: Anaphora Resolution, Longman, Londýn 2002, str. 14

1.3 Komplikované druhy anafor

Nepřímá anafora

Nepřímá anafora je nepřímý odkaz k antecedentu záležící na čtenářovy/posluchačovy znalosti světa.

- (1.1) Když *Take That* rozpadla, kritici nedali *Robbie Williamsovi* žádnou šanci na úspěch.⁴

Robbie Williams byl členem bývalé britské popové skupiny *Take That*, a tak čtenář by to musel vědět, aby to mohl vnímat jako nepřímý odkaz.

Tento vztah je charakterizován také jako vztah části něčeho k něčemu jako celku nebo vztah členu množiny k množině.

- (1.2) Po včerejším tréninku mě bolí celé *tělo*, nejvíc *obě nohy*.

Referenční identita a smyslová identita

Referenční identita (identity-of-reference) je případ anafory, kdy anafor a jeho antecedent ukazují na jednu a tu samou věc v reálném světě, přitom smyslová identita (identity-of-sense) nastává, když anafor a antecedent pouze ukazují na věci, které mají stejný popis ([Mitkov, 2001]).

- (1.3) Jan koupil *koláč* a hned *ho* snědl. (referenční identita)

- (1.4) Jan koupil *koláč*, Marie koupila také jeden \emptyset . (smyslová identita)

Druhy antecedentů

Antecedentem může být nejenom jmennou frází, ale také koordinovanými jmennými frázemi (dvě nebo více jmenných fráz jsou koordinovány čárkou, spojkou), slovesem, slovesnou frází, větou, posloupností vět.

⁴ Mitkov, R.: *Anaphora Resolution*, Longman, Londýn 2002, str. 15

- (1.5) *Nejvyspělejší země světa (USA, SRN, Japonsko aj.), které vyrostly na konzervativních hodnotách a Ø hájí je, vystavují veškeré poznání okamžitě experimentu.*

Nejednoznačná anafora

- (1.6) Franta řekl Janovi, že Ø je asi v nebezpečí.

V tomhle případě je anafora nejednoznačná – nulové zájmeno může odkazovat jak k Frantovi, tak k Janovi. Mnohdy nejednoznačnost závisí na sémantice slovesa nebo na jiných členech věty nebo diskurzu. V následujícím příkladu je anafora podle sémantiky řídicího slovesa jednoznačně určena:

- (1.7) Franta varoval *Jana*, že Ø je asi v nebezpečí.

Kapitola 2

Analýza anafory

2.1 Předpoklady k určení anafory

K analýze anaforů je potřeba znát od morfologických k sémantickým a pragmatickým pravidlům.

Morfologické předpoklady

Na základě skutečnosti, že nominální anafory se často shodují se svými antecedenty v rodě a čísle, lze někdy jednoznačně pro ně určit antecedent.

(2.1) *Marie* dostala od kamarádů dárek. Ø Slavila 20. narozeniny.

Ze tří substantiv {Marie, kamarádi, dárek} jen *Marie* splňuje podmínku shody v rodě a čísle s nulovým anaforem *ona*.

Syntaktické předpoklady

Syntaxe hraje důležitou roli při hledání antecedentu z možných kandidátů. Například reflexivní anafor odkazuje téměř vždy k subjektu v té samé větě:

(2.2) *Jana* půjčila *své* kamarádce *svoji* oblíbenou knihu.

Sémantické předpoklady

Ve větě:

(2.3) *Kočka* vylezla na střechu. Ø Opalovala se tam.

by pravidlo shodování v rodě a čísle nestačilo. Sloveso *opalovat se* vyžaduje životného aktora, takže sémantická informace o životnosti *kočky* by byla nápomocná. Zvlášť v tomto případě by stačily prioritní syntaktické tabulky od Lappina & Leasse, podle

keré by *kočka* dostala větší prioritu nad *střechu*, protože *kočka* je subjekt a *střecha* příslovečné určení místa.

Diskurzni předpoklady

Další důležitou roli v analýze anafory hrají diskurzni znalosti – znalosti **aktuálního členění věty**. Jako aktuální členění věty se označuje členění věty na **základ** (východisko, téma, topic) a **jádro** (ohnisko, réma, focus). Podle aktuálního členění je český slovosled zpravidla vázán výpovědní dynamičností: věta začíná vlastním tématem a jeho částmi, pak následují části rématu – nejdřív sloveso, pak další části a nakonec vlastní ohnisko (nositel intonačního centra, nejdynamičtější člen věty). K charakteristice všech výskytů slov bez ohledu na jejich okolí slouží pojem kontextového zapojení. Slovo ve větě je kontextově zapojené (často ve východisku), jestliže je nositelem informace „dané“, a naopak slovo je kontextově nezapojené (často v jádře), nese-li informaci „novou“.

- (2.4) Od té doby pobíral *Lojzík* invalidní rentu, \emptyset pomáhal tátovi v pekárně, a dokonce \emptyset se i oženil; jenom někdy \emptyset pozoroval, že na tu nohu, co *mu* ji Oberhuber upřel, jakoby drobet \emptyset kulhá nebo \emptyset napadá; ale i tomu \emptyset byl rád, že to aspoň vypadá, jako by \emptyset měl protézu.

V případech nejednoznačnosti antecedentu {*Lojzík*, *táta*, *Oberhuber*}, vyhraje ten kandidát, který má nejvyšší stupeň aktivovanosti (salience) – *Lojzík*, k němuž odkazuje posloupnost nulových zájmen a zájmeno *mu*. V [Hajičová et al., 2001] a [Hajičová, 2003] jsou navržena základní pravidla určující stupně aktivovanosti.

Předpoklady znalosti světa

- (2.5) Máme doma *kanárka*. Když tatínek ráno vstává, \emptyset vyskočí na bidýlko, \emptyset rozhlédne se a \emptyset začne zpívat.

Kdyby systém analýzy anafory neměl informaci, že činnost „vyskočení na bidýlko“ přísluší spíš *kanárkovi* než *tatínkovi*, těžko by s tímhle případem poradil. Pomůže mu i znalosti aktuálního členění věty.

Následující věty jsou dalším pěkným příkladem poukazujícím na důležitost znalosti světa:

- (2.6) a. Vojáci stříleli na muže. Padli.
b. Vojáci stříleli na muže. Netrefili.⁵

2.2 Známé způsoby analýzy anafory

Počáteční pokusy o rozluštění anaforů v 60. a 70. letech byly založeny na heuristických pravidlech a nevyužívaly maximálně lingvistické rozborů. Na konci 70. let se objevila první práce zaměřená na diskurz a na konci 80. let se dokonce pracovalo už s některými znalostmi světa. Dále budou stručně popsány některé z nejvýznamnějších prací na anafoře.

2.2.1 Hobbsův algoritmus se stromovým vyhledáváním (1978)

Hobbsův algoritmus používá syntaktické zobrazení vět a vykonává hledání jmennofrázového antecedentu na těchto syntaktických stromech. K vytváření stromů vybral Hobbs nekontextovou gramatiku. Jednotlivé kroky algoritmu zajišťují, aby jmenná fráze měla odpovídající rod a číslo jako anafor a aby nereflexivní zájmeno a jeho antecedent nebyly ve stejné větě. Umějí se poradit i s kataforou. Jako antecedent k zájmenu *they* vybírají sémanticky slučitelné prvky.

Po manuálním vyhodnocení svého algoritmu na 300 zájmenech ze tří žánrově různých textů vyšla Hobbsovi míra úspěšnosti 88,3% (91,7% pro vylepšený algoritmus dalšími vylučujícími faktory).

$$\text{Míra úspěšnosti} = \text{Recall} = \frac{\text{Počet správně vyřešených anaforů}}{\text{Celkový počet anaforů}}$$
$$\text{Precision} = \frac{\text{Počet správně vyřešených anaforů}}{\text{Počet vyřešených anaforů}}$$

Obr. 2.1: Míra úspěšnosti, recall a precision pro algoritmus analýzy anafory

⁵ Mitkov, R.: *Anaphora Resolution*, Longman, Londýn 2002, str. 15

2.2.2 BFP algoritmus založené na teorii centrování (1987)

BFP algoritmus od Brennana, Friedmana a Pollarda byl rozšířen z teorie **centrování** od Grosze (1986). Centrování je teorie o diskurzni koherenci a je založeno na myšlence, že každá výpověď obsahuje tématicky význačný prvek nazvaný **centrum**. Každé výpovědi je přiřazena množina potenciálních dalších centrů, která korespondují diskurzním prvkům ve výpovědi a jedno centrum, které je aktuální pro každou výpověď. BFP algoritmus generuje možné kombinace aktuálního a potenciálního centra, z nichž pak vybere nejpravděpodobnější pár podle stupní spojitosti výpovědi.

Podobně jako Hobbsův algoritmus, byl vyvíjen na základě předpokladu, že na vstupu jsou správné syntaktické struktury. K tomu, aby podstoupil automatické ohodnocení na přirozeně vyskytujících datech, musel by BFP algoritmus být specifikován do větších detailů.

Walker později provedl ruční ohodnocení BFP a Hobbsova algoritmu na korpusu s 281 příklady. BFP dosáhl míry úspěšnosti 77,6%, Hobbsův 81,8%.

2.2.3 Lappin & Leassův algoritmus se systémem vah (1994)

Lappin & Leass popisuje algoritmus s jednoduchým systémem vah pro zájmena 3. osoby (včetně reflexivní a reciprokální). Algoritmus se spočívá na váhových mírách odvozených ze syntaktické struktury a na jednoduchém dynamickém modelu, který vybírá antecedent zájmena ze seznamu kandidátů. Kandidáty na antecedent zájmena jsou jmenné fráze seřazeny podle následující hierarchie:

subjekt > agens (konatel děje ve větách s přísudkem v trpném rodě) > objekt > nepřímý objekt nebo obligátní příslovečné určení (obsaženo ve slovesné valenci) > fakultativní příslovečné určení.

Váhu navíc dostanou:

- jmenné fráze v aktuálně zkoumané větě
- jmenné fráze, které nejsou obsaženy v jiné jmenné frázi
- jmenné fráze, které nejsou v příslovečně-předložkových frázích

Lappin & Leassův algoritmus vůbec nepoužívá sémantické informace nebo znalosti světa. Tento algoritmus byl ohodnocen na korpusu počítačových manuálů o obsahu přibližně 82.000 slov a uspěl na 86%.

2.2.4 Mitkovův robustní algoritmus, který nepotřebuje mnoho znalosti světa

Mitkovým cílem je vytvořit systém pro analýzu zájmen, který by byl rychlý, cenově dostupný a spolehlivým alternativem systému založeného na znalostech. ([Mitkov]) Jeho algoritmus se vyhýbá komplexním syntaktickým, sémantickým a diskurzním analýzám. Místo toho se spoléhá na seznamu preferencí, zvaný indikátory antecedentu. Pro každý anafor vyhledá algoritmus jmenné fráze ve větě, kde je anafor, a v předchozí větě; aplikuje indikátory (které přičítá nebo odečítá body) na tyto kandidáty, a jmenná fráze s nejvyšším výsledkem je navržen jako antecedent.

Body navíc dostávají:

- jmenné fráze, které se opakují víckrát v paragrafu, kde se nachází zájmeno
- jmenné fráze, které mají identický kolokační vzor jako zájmeno

Penalizovány jsou neurčité jmenné fráze a jmenné fráze v příslovečném určení. Kromě toho Mitkov předdefinuje seznam sloves a řadu dalších vzorových větných struktur. Jmenným frázím, které se objeví hned po slovesu ze seznamu nebo v těchto větných strukturách budou přčteny další body. Pro úplný výčet a detailnější popis indikátorů viz [Mitkov, 2001].

Mitkovův algoritmus uspěl na 223 anaforických zájmenech z různých uživatelských manuálů na 89,7%. Algoritmus byl aplikován i na arabštině, polštině, francouzštině a bulharštině. V bulharských turistických textech byl MARS (Mitkov's Anaphora Resolution System) správně ohodnocen na 68,1%.

2.3 Strojové učení jako nástroj pro analýzu anafory

Termín strojové učení je často používán ke specifickému označení metod, které reprezentují naučené znalosti v deklaraticní, symbolické formě jako protiklad k více numericky-orientovaným statistickým nebo neuronu-síťovým trénovacím metodám.

Obsahuje zejména metody, které reprezentují naučené znalosti ve formě interpretovatelných rozhodovacích stromů, logických pravidel a instancí ([Mooney, 2002]). Jako dalším alternativem systému založeného na znalostech nabízí strojové učení nabytí těchto znalostí z množin příkladů z anotovaného nebo neanotovaného korpusu.

Rozhodovací stromy jsou klasifikační funkce reprezentovány jako stromy, ve kterých jsou uzly atributy testů, hrany (větve) hodnoty atributů a listy třídy. Algoritmy s rozhodovacími stromy ID3 a C4.5 od Quinlana ([Quinlan, 1993]) jsou v analýze anafory používány nejčastěji.

2.3.1 Řešení od Aone a Bennetta

Aone a Bennett popisuje systém analýzy anafory pro japonštinu, který je trénovaný na korpusu novinových článků anotovaném diskurzními informacemi (organizace, lidé, polohy...). Systém používá rozhodovací stromy C4.5. Trénovací korpus obsahoval 1971 anaforů. Výsledek je: recall = 67,53% (70,2% bez určení typů anaforů); precision = 76,27% (72,27%).

2.3.2 Řešení od McCarthyho a Lehnerta

Systém od McCarthyho a Lehnerta používá rozhodovací stromy C4.5 na korpusu MUC-5 English Joint Venture a řeší pouze specifické druhy jmenných frází (organizace a obchodní entity). V tomto omezeném žánru a navíc na korpusu, kde chyby byly ručně opraveny před procesem, uspěl systém s: recall = 85,4% (98,1% s prořezáváním); precision = 87,6% (92,4%).

2.3.3 Řešení od Soon, Ng a Lima

Narozdíl od omezeného systému na japonštině od Aone a Bennetta a od omezeného systému na obchodní prvky od McCarthyho a Lehnerta navrhuje Soon, Ng a Lim řešení pro všechny druhy jmenných frází na neomezeném žánru textů. Využívá k tomu C4.5 algoritmus, korpusy MUC-6 a MUC-7 a WordNet (pro určení sémantických tříd). Trénovací data sčítá kolem 13.000 slov, data na ohodnocení má velikost 14.000 slov. Úspěšnost: recall = 52%, precision = 68%.

2.4 Další způsoby analýzy anafory

Kromě výše popsaných způsobů analýzy anafory existuje ještě mnoho jiných řešení. Některé jsou vylepšené verze z dosud známých algoritmů (Kennedy & Boguraev), s přidanými genetickými algoritmy pro optimalizaci (Orasan a kol. 2000, Mitkov a kol. 2002). Některé používají jinou metodu, např. pravděpodobnostní (Ge, Hale a Charniak), nebo jinou teorii, např. teorii binding, teorii focusing nebo teorii diskurzí reprezentace. Některé se snaží využívat maximálně všeobecné dostupné znalosti světa (Rich a LuperFoy, Carbonell a Brown), ale jsou náročné na realizaci a vyžadují značně velký lidský potenciál.

Kapitola 3

Anafora v Pražském závislostním korpusu

Korpusy hrají důležitou roli v analýze anafory. Přinášejí prospěch pro trénování a vyhodnocení algoritmů analýzy anafory, zvláště když jejich anotace překrývá nejenom jednotlivé páry anafor-antecedent, ale také koreferenční řetězce. Pro analýzu anafory v českém jazyce splňuje tyto podmínky Pražský závislostní korpus.

3.1 Pražský závislostní korpus

Pražský závislostní korpus (PDT – The Prague Dependency Treebank) je projekt na automatickou a manuální anotaci podstatného množství dat českého jazyka lingvisticky s bohatými informacemi od morfologie a syntaxi až k sémantice/pragmatice a dále. PDT obsahuje velké množství českých textů se vzájemně propojenými morfologickými (2 milióny slov), syntaktickými (1,5 MS) a složenými podloženými syntaktickými a sémantickými anotacemi (0,8 MS).

PDT je založen na dlouholeté Pražské lingvistické tradici a je přizpůsoben k současným potřebám počítačového lingvistického výzkumu. Cílem PDT je mapovat teoretické úspěchy Pražské lingvistické školy do reálních jazykových dat a umožnit použití metod strojového učení k automatickým analýzám s rozumnou přesností. V PDT je na rozsáhlý jazykový materiál aplikován **Funkční generativní popis** (FGP).

FGP je popis jazyka založený na závislostní syntaxi, kde je sloveso chápáno jako centrum věty, jejíž jiné členy se považují za závislé na slovesu. Orientuje se na formální popis jazyka, mající matematické vlastnosti a jasně vymezená kritéria pro klasifikaci jednotlivých jevů. Pracuje s několika rovinami popisu, tzv. stratifikační popis.

Data v PDT jsou anotována na třech rovinách:

- morfologické rovině
- analytické rovině
- tektogramatické rovině

Všechny roviny byly automaticky předzpracovány, a pak ručně opravovány a doplněny dalšími informacemi anotátory.

3.1.1 Morfologická rovina

Anotace (značkování) každé věty se skládá z posloupnosti lematizovaných a tagovaných značek. Lema je základní nebo normalizovaný tvar slova a shoduje se se slovníkovým tvarem. Tag je morfologická sestava obsahující 15 znaků, která vyjadřuje slovní druh a různé morfologické kategorie. Tag byl vyvinut na Ústavu formální a aplikované lingvistiky a je používán také v existujícím morfologickém slovníku češtiny.

3.1.2 Analytická rovina

Analytická rovina zachycuje věty v jejich povrchové podobě. Věta je reprezentována jako uspořádaný kořenový strom opatřený uzly a hranami. Každé jednotka z morfologické rovinami je reprezentována přesně jedním uzlem stromu a závislostní vztah mezi dvěma uzly je zachycen hranou mezi nimi. Některé hrany charakterizují také různé lingvistické nebo technické jevy (např. koordinace, apozice, interpunkce atd.) Lineární seřazení uzlů, které koresponduje originálnímu větnému slovosledu, je také zaznamenáno.

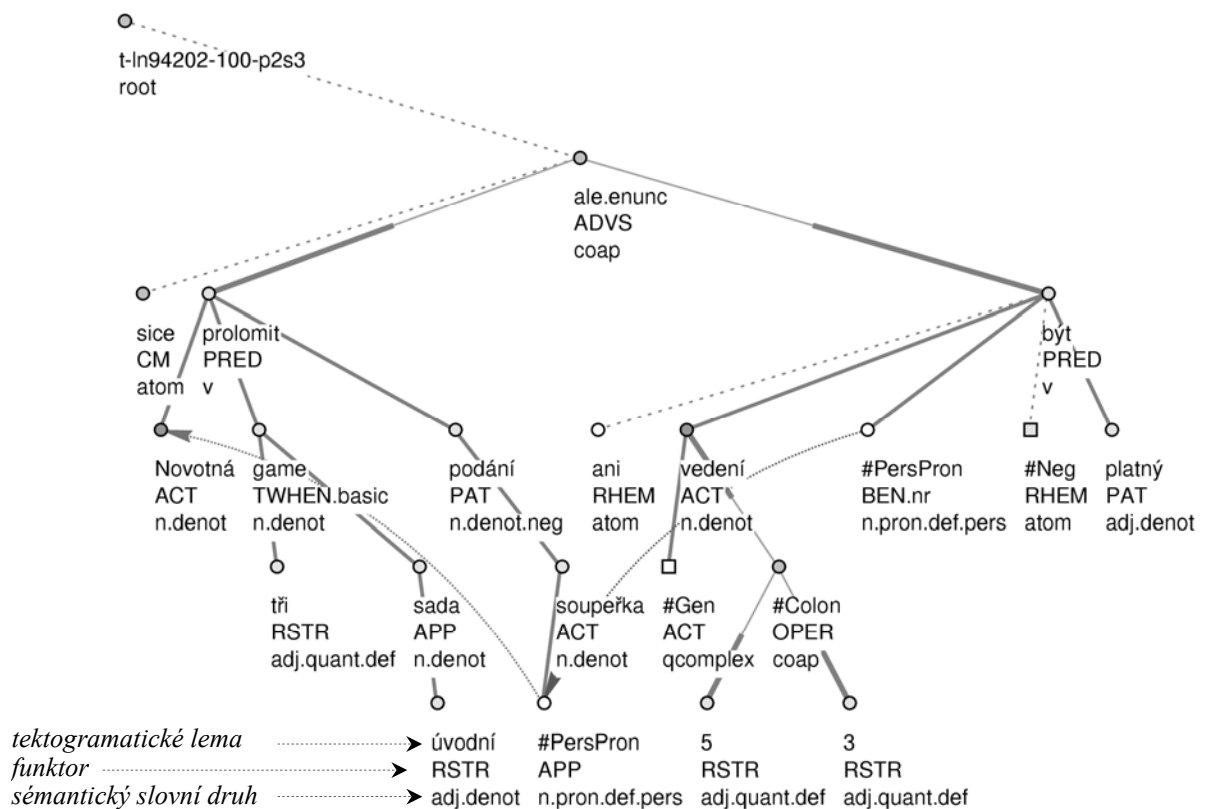
3.1.3 Tektogramatická rovina

Tektogramatická reprezentace věty zachycuje následující aspekty:

- **tektogramatická strukturu a funkory pomocí stromové struktury.** Každý strom představuje hloubkovou strukturu věty. Hodnotami uzlů ve stromu jsou jen autosémantická (významová) slova (s výjimkou některých technických uzlů). Koreláty synsématických (pomocných) slov jsou připojeny k uzlům s autosémantickým slovem, tedy pomocná slovesa a podřadící spojky ke slovesům, předložky k substantivům ap. V případech vypouštění na povrchové rovině (elidovaných členů věty) se do tektogramatického stromu doplňují další uzly pro vypouštěná slova.

- **aktuální členění.** Každý uzel má jednu ze tří hodnot, vyjadřujících kontextové zapojení:
 - uzel je kontextově zapojený
 - uzel je kontrastně, opačně zapojený
 - uzel není kontextově zapojený
- **koreference** (viz kapitola 3.2)

Zápis věty na tektogramatické rovině je zbaven homonymie (víceznačnosti). Oproti analytické rovině splňují závislostní stromy na tektogramatické rovině podmínku projektivity: pořadí uzlů (jejich lineární uspořádání) zachovává hloubkový slovosled. Ten odpovídá tzv. výpovědní dynamičnosti, a spolu s určením základu a jádra udává aktuální členění věty.



Obr. 3.1: Příklad věty na tektogramatické rovině: *Novotná sice prolomila ve třetím gamu úvodní sady podání své soupeřky, ale ani vedení 5:3 jí nebylo platné.* (In94202_100#5)

Tektogramatické lema je jedním z atributů uzlu tektogramatického stromu. Má podobu řetězce grafémů základní slovní formy, když uzel reprezentuje slovo přítomné

v povrchové podobě věty, nebo obsahuje hodnotu „umělou“, pokud je uzel nově vytvořený.

#PersPron	osobní, posesivní a reflexivní zájmena	<i>přítomna v povrchové podobě věty i nově vytvořena</i>
#Cor	nevyjádřený člen závislého slovesa kontroly	<i>nově vytvořený</i>
#QCor	nevyjádřený člen závislého substantiva kontroly	<i>nově vytvořený</i>
#Gen	nepřítomný všeobecný aktant	<i>nově vytvořený</i>
#Rcp	nevyjádřené valenční doplnění z důvodu reciprokalizace	<i>nově vytvořené</i>
#Benef	nevyjádřené volné doplnění s významem benefaktoru v konstrukcích s kontrolou	<i>nově vytvořené</i>

Tab. 3.1: Seznam zástupných tektogramatických lemat (uměle vytvořených lemat) používaných v diplomové práci

Jedna z informací, která obsahuje uzel na tektogramatické rovině, jsou funktoři. **Funktoři** chápeme jako sémantické ohodnocení syntaktického vztahu závislosti, jako funkce doplnění (lexikálních jednotek) ve větné struktuře.⁶ Uvedeme zde popis těch funktořů⁷, které se objevují v naší práci.

ACT	aktant – aktor	<i><u>Otec pracuje.</u></i>
ADDR	aktant – adresát	<i>Poslal dárek <u>příteli.</u></i>
APP	volné doplnění substantiva vyjadřující přináležitost	<i>můj <u>hrad</u></i>
AUTH	volné doplnění substantiva označující autora	<i><u>Nezvalovy verše</u></i>
BEN	volné doplnění vyjadřující ne/prospěch	<i>Pracuje <u>pro firmu.</u></i>
COMPL	volné doplnění – doplněk	<i>Vrátila se <u>unavená.</u></i>
DENOM	efektivní kořen nezávislé nominativní klauze, která není vsuvkou	<i>Základní <u>škola.</u></i>
EFF	aktant – efekt	<i>Jmenovali ho <u>předsedou.</u></i>
LOC	volné doplnění místa odpovídající na otázku „kde“	<i>Pracuje <u>v Praze.</u></i>
ORIG	aktant – origo	<i>Vyrábí nábytek <u>ze dřeva.</u></i>

⁶ V [PDT-manuál] str. 423

⁷ Popis a příklady byly vzaty z manuálu [PDT-manuál], str. 425-427

PAT	aktant – patiens	<i>Vaří <u>oběd</u>.</i>
PRED	efektivní kořen nezávislé slovesné klauze, která není vsuvkou	<i>Pavel <u>dal</u> kytku Martině.</i>
RSTR	volné doplnění blíže specifikující řídicí substantivum/efektivní kořen vztažné klauze	<i><u>velký</u> dům/Udeřil i toho, kdo si to <u>nezasloužil</u>.</i>

Tab. 3.2: Seznam funktořů používaných v diplomové práci

Sémantické slovní druhy jsou kategoriemi tektogramatické roviny, odpovídají základním onomaziologickým kategoriím (substance, vlastnost, okolnost, událost) a nejsou totožné s deseti „tradičními“ slovními druhy. Sémantické slovní druhy jsou čtyři: sémantická substantiva, sémantická adjektiva, sémantická adverbia a sémantická slovesa. Pro naši práci bude potřeba porozumět sémantickým substantivům, protože bývají často antecedenty, proto je jim věnována následující tabulka.

n.denot	pojmenovací sémantické substantivum
n.denot.neg	pojmenovací sémantické substantivum s odděleně reprezentovaným příznakem negace
n.pron.def.demon	určité pronominální sémantické substantivum ukazovací
n.pron.def.pers	určité pronominální sémantické substantivum osobní
n.pron.indef	neurčité pronominální sémantické substantivum
n.quant.def	určité kvantifikační sémantické substantivum

Tab. 3.3: Sémantická substantiva

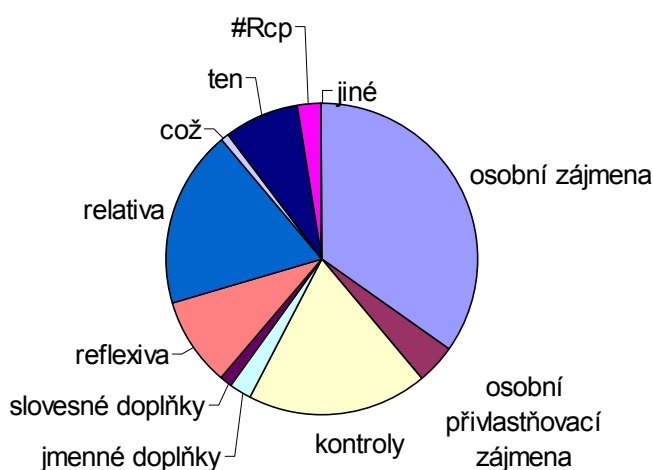
3.1.4 Data v PDT

Data v PDT jsou celé anotované články z následujících novin a časopisů: Lidové noviny, Mladá fronta Dnes, Českomoravský Profit, Vesmír. Texty v elektronické podobě poskytnul institut Českého národního korpusu. Anotace jednotlivých rovin nepokrývají data rovnoměrně. Čím složitější je rovina, tím méně dat je na ní anotováno, neboť anotace na složitější rovině potřebuje více času, zdrojů a lidské práce. Všechny soubory anotovány na vyšší rovině jsou anotovány také na nižších rovinách.

Data jsou rozdělena do tří skupin: trénovací data (~ 80%), testovací na rozvoj dat (~ 10%) a testovací data na hodnocení (~ 10%).

3.2 Koreference v PDT

V PDT se koreference rozděluje na dva druhy: gramatickou a textovou. Gramatická koreference je realizována zvratnými zájmeny, vztažnými prostředky (vztažná zájmena, zájmenné příslovce, spojovací výraz *což*), doplňky (vyjádřené slovesem nebo jménem), recipitou (#Rcp) a kontrolou (nevyjádřeným subjektem infinitivu). Textová koreference používá osobní zájmena v 3. osobě a ukazovací zájmena (zájmeno *ten*).



Obr. 3.2: Relativní četnost koreferujících výrazů v trénovacích datech PDT (celkem 36.852 anaforů)

V práci [Kučová et al., 2003] bylo představeno několik ručně psaných pravidel:

- pravidlo na určení anaforů s precision 98,3%
- pravidlo na určení antecedentů relativ s precision 95,43%
- pravidlo na určení antecedentů reflexiv s precision 87,25%
- pravidlo na určení antecedentů kontroly s precision 88,64% pro kontrolu ADDR, 69,93% pro kontrolu ACT, 33,33% pro kontrolu PAT (více o skupinách kontroly v následující kapitole)

Kučová a Žabokrtský ve své práci [Kučová et Žabokrtský, 2005] navrhl soubor pravidel pro řešení osobní zájmenné anafory s mírou úspěšností 60,4%. Používá postupně různé filtry k odstranění nepravděpodobných kandidátů, které jsou z předchozí věty nebo ze stejné věty jako osobní zájmeno:

1. Kandidáti, kteří jsou ze stejné věty jako anafor a nepředchází ho, jsou odstraněni.
2. Kandidáti, kteří nejsou sémantickými substantivy, jsou odstraněni.
3. Kandidáti, kteří jsou subjekty ve stejné klauzi jako anafor, jsou odstraněni.
4. Kandidáti, kteří se neshodují s anaforem v čísle nebo v rodě, jsou odstraněni.
5. Kandidáti, kteří jsou rodičem nebo prarodičem anafora ve stromě, jsou odstraněni.
6. Je-li v seznamu kandidátů dvojice uzel a jeho potomek, pak je potomek odstraněn.
7. Je-li v seznamu kandidátů kandidát se stejným funktoem jako anafor, pak jsou všichni kandidáti s odlišným funktoem odstraněni.
8. Je-li v seznamu kandidátů kandidát v pozici subjektu, pak jsou všichni kandidáti v jiné pozici než subjekt odstraněni.

Na konci je kandidát, který je nejbližší k anaforu, vybrán jako antecedent.

Kapitola 4

Analyzátor české anafory

Analyzátor české anafory (AČA) se zabývá určením antecedentů u osobní zájmené anafory, nulové anafory kontroly, reflexivní zájmené anafory, relativní zájmené anafory, demonstrativní zájmené anafory, nulové anafory reciprocity, a navíc i jmennou závilostí u doplňků. Používá k tomu systém strojového učení C4.5 a ručně psaná pravidla; pro přístup k závislostním stromům a jejich uzlům v PDT používá program btred. K trénování a testování algoritmů slouží trénovací a testovací data z tektogramatické roviny v PDT.

AČA se nevěnuje problémům automatického rozeznávání anaforů, ty byly úspěšně vyřešeny v práci (Kučová) na 98,3%.

4.1 Osobní zájmená anafora

AČA zkoušel vyřešit problém osobní a posesivní zájmené anafory pomocí systému strojového učení C4.5 nebo pomocí ručně psaných pravidel.

4.1.1 Řešení pomocí strojového učení

AČA vytvoří pro každé osobní zájmeno seznam kandidátů tvořený ze sémantických substantiv, nacházejících ve stejné větě jako anafora nebo v předcházející větě. Sémantická substantiva prošla testem shody v čísle a rodě, který se také ohlíží na případy složeného antecedentu, kdy se antecedent skládá z více substantiv.

Pro třídy osobní zájmená anafora, posesivní zájmená anafora a non-anafora sestavil AČA vektor 18 atributů určených z dvojic anafora a kandidáta.

Při sestavení atributů pro osobní zájmenou anaforu jsme vycházeli z toho, že osobní zájmena zastupují názvy osob, zvířat a věci, proto se shodují se svým antecedentem v mnoha vlastnostech a funkcích. Dále jsme použili poznatky z Lappin & Leassova,

Mitkovova algoritmu a teorie aktuálního členění věty. Také jsme chtěli využít co nejvíc informací, které obsahují uzly anafora a kandidáta na tektogramatické a analytické rovině.

- *závislost kandidáta*: Možné hodnoty {PRED, V, N, other, none}. Hodnota PRED platí pro případ, kdy kandidát závisí na predikátu věty; V – závislost na slovese; N – závislost na substantivu; other – závislost na něčem jiném; none – kandidát je kořenem stromu.
- *závislost anaforu*: Možné hodnoty {PRED, V, N, other, none}. Hodnota PRED platí pro případ, kdy anafor závisí na predikátu věty; V – závislost na slovese; N – závislost na substantivu; other – závislost na něčem jiném; none – kandidát je kořenem stromu.
- *shoda v závislosti*: Možné hodnoty {ano, ne}. Závislost kandidáta = závislost anaforu: ano/ne.
- *funktor kandidáta*: Možné hodnoty {ACT, AUTH, PAT, ADDR,...}.
- *funktor anaforu*: Možné hodnoty {ACT, AUTH, PAT, ADDR,...}.
- *shoda ve funktoru*: Možné hodnoty {ano, ne}. Funktor kandidáta = funktor anaforu: ano/ne.
- *kandidát je aktant*: Možné hodnoty {ano, ne}. Je-li funktor kandidáta ACT/ADDR/PAT/EFF/ORIG, hodnota je pozitivní; jinak je negativní.
- *anafor je aktant*: Možné hodnoty {ano, ne}. Je-li funktor anaforu ACT/ADDR/PAT/EFF/ORIG, hodnota je pozitivní; jinak je negativní.
- *oba jsou aktanty*: Možné hodnoty {ano, ne}. Kandidát je aktant & anafor je aktant: ano/ne.
- *kandidát je subjekt*: Možné hodnoty {ano, ne}. Je-li kandidát subjektem věty, hodnota je pozitivní; jinak je negativní.
- *anafor je subjekt*: Možné hodnoty {ano, ne}. Je-li anafor subjektem věty, hodnota je pozitivní; jinak je negativní.
- *oba jsou subjekty*: Možné hodnoty {ano, ne}. Kandidát je subjekt & anafor je subjekt: ano/ne.
- *vzdálenost*: Možné hodnoty {0, 1, 2, 3,...}. Vzdálenost anaforu od kandidáta v textu: 0 – anafor a kandidát jsou ve stejné větě; 1 – kandidát je v předchozí větě atd.

- *ve stejné klauzi*: Možné hodnoty {ano, ne}. Anafor a kandidát jsou ve stejné klauzi: ano/ne.
- *aktuální členění kandidáta*: Možné hodnoty {t, f, c}; t – kandidát je kontextově (nekontrastivně) zapojený výraz; f – kandidát je kontextově nezapojený výraz; c – kandidát je kontextově kontrastivně zapojený výraz.
- *aktuální členění anaforu*: Možné hodnoty {t, f, c}; t – anafor je kontextově (nekontrastivně) zapojený výraz; f – anafor je kontextově nezapojený výraz; c – anafor je kontextově kontrastivně zapojený výraz.
- *shoda v aktuálním členění*: Možné hodnoty {ano, ne}. Aktuální členění kandidáta = aktuální členění anaforu: ano/ne.
- *kandidát je frekventovaný*: Možné hodnoty {ano, ne}. Je-li výskyt kandidáta v souboru víc než jednou, hodnota je pozitivní; jinak je negativní.

4.1.2 Řešení pomocí ručně psaných pravidel

Ručně psaná pravidla AČA jsou hlavně inspirovány Lappin & Leassovým a Mitkovovým algoritmem. Jsou vůči nim pozměněny kvůli rozdílu v žánru dat (Lappin & Leassův a Mitkovův algoritmus byl vyvíjen a testován na odborných manuálech) a v jazyce. Dalším výrazným rozdílem je struktura dat. Žádný ze dvou zmíněných nemá data v závislostních stromech.

Postup

Sémantická substantiva v předcházející a ve stejné větě jako anafor jsou testována shodou v čísle a rodě. Všem substantivům, která prošla filtrem, jsou pak přiřazeny váhy, pozitivní váha naznačuje pravděpodobnost, že dané substantivum je antecedent. Naopak negativní váha signalizuje nedostačující jistotu, že je to antecedent současného zájmena. Ty váhy jsou:

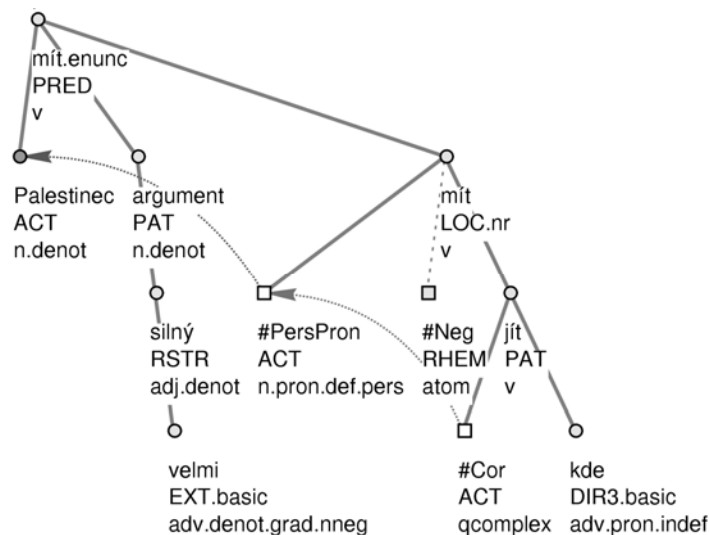
- *subjekt*: Váha +1 je přiřazena subjektu ve větě.
- *subjekt v hlavní větě*: Váha +1 je přiřazena subjektu v hlavní větě.
- *frekventované substantivum*: Váha +1 je přiřazena těm substantivům, která se vyskytují v souboru víc než jeden krát.
- *nejčastější funktor*: Váha +1 je přiřazena substantivům, která mají funktor ACT/ADDR/PAT/APP (viz obr. 4.1) a závisí na slovese.

- *negace nejčastějšího funktoru*: Váha -1 je přiřazena substantivům, která nemají funktor ACT/ADDR/PAT/APP nebo nezávisí přímo na slovese.
- *kolokace*: Váha +2 je přiřazena substantivům, která mají stejnou kolokaci, jakou má zájmeno.
- *vzdálenost*: Váha +2 je přiřazena substantivům vyskytujícím ve stejné větě jako anafor a předcházejí ho; váha +1 je přiřazena substantivům ve větě předcházející větu s anaforem.

Kandidát s nejvyšší váhou je pak vybrán jako antecedent zájmena. Je-li kandidátů s nejvyšší váhou víc, AČA vybere ten, který je v textu nejbliž vlevo od anaforu. Jsou-li všichni kandidáti vpravo od zájmena, antecedent bude ten nejbliž vpravo od zájmena.

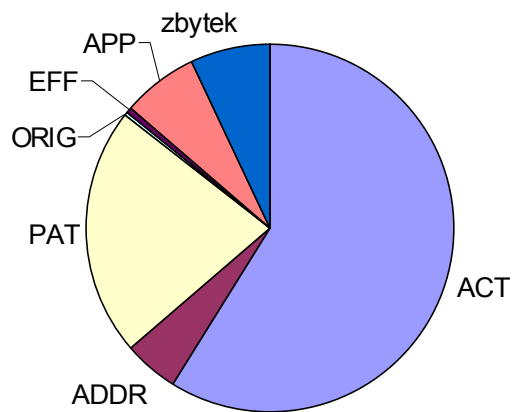
Seznam kolokací je vytvořen ze sloves a jejich seznam pojmenovacích sémantických substantiv, která se objevila v textu jako jejich aktanty. Ukážeme to na následujícím příkladu:

- (4.1) Je to tragédie, srážka mezi dvěma spravedlivými. Palestinci mají velmi silný argument v tom, že \emptyset nemají kam jít;



Obr. 4.1: Ilustrační příklad: *Palestinci mají velmi silný argument v tom, že nemají kam jít;* (ln94207_92#48)

Sloveso *mít* má kolokaci {Palestinec-ACT}, a tak při hledání antecedentu pro \emptyset -ACT, které také závisí na slovesu *mít*, dostane kandidát *Palestinec* váhu +2 za shodu kolokace.

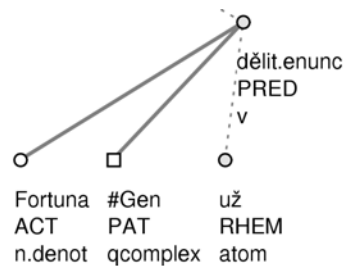


Obr. 4.2: Relativní četnost funktorů antecedentů osobní zájmené anafory v trénovacích datech PDT

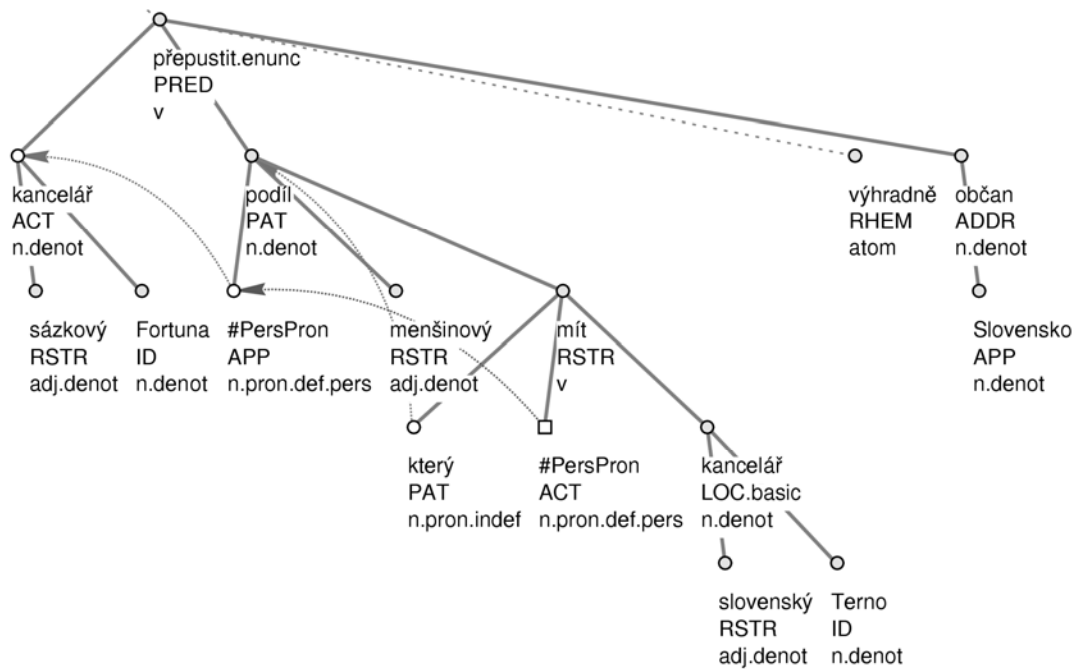
Ilustrace

(4.2) Fortuna_i už „dělila“

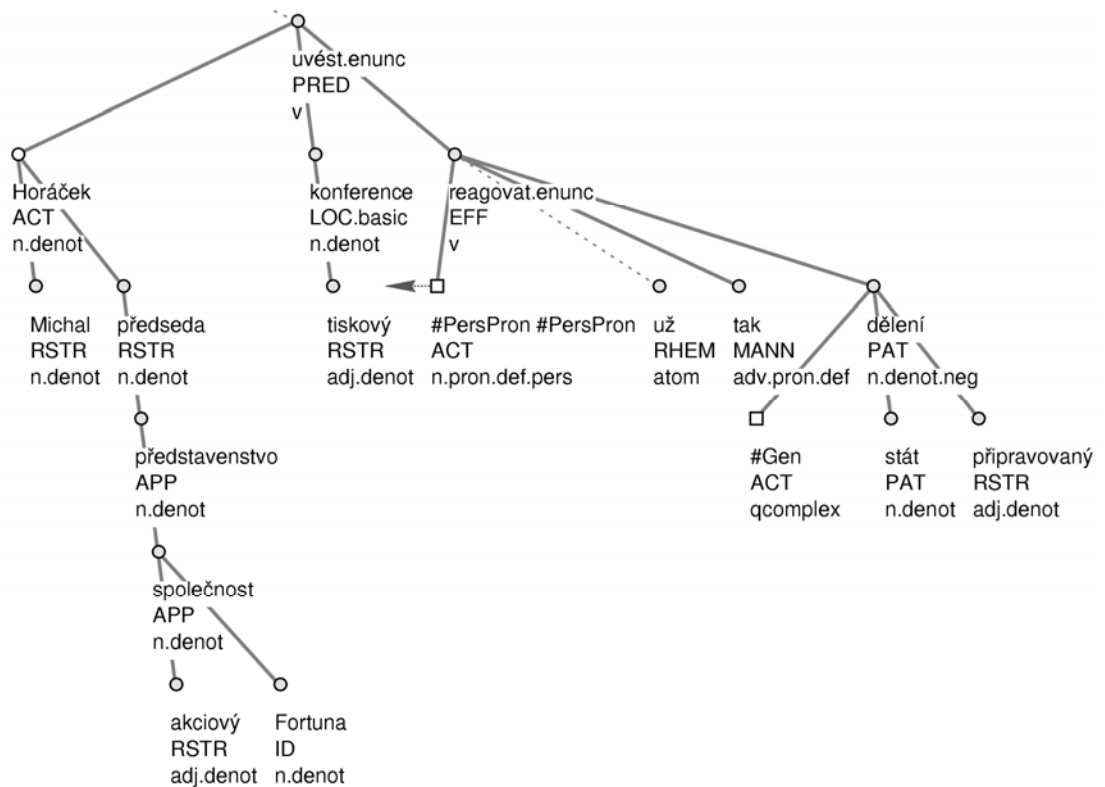
Sázková kancelář_i Fortuna_j přepustila svůj menšinový podíl, který Ø_i měla ve slovenské kanceláři_j Terno, výhradně občanům Slovenska. Ø_j Reagovala už tak na připravované dělení státu, uvedl na tiskové konferenci předseda představenstva a. s. Fortuna_k Michal Horáček.



Obr. 4.3: Ilustrační příklad: *Fortuna už „dělila“* (mf920922_074#1)



Obr. 4.4: Ilustrační příklad: *Sázková kancelář Fortuna přepustila svůj menšinový podíl, který měla ve slovenské kanceláři Terno, výhradně občanům Slovenska.* (mf920922_074#2)



Obr. 4.5: Ilustrační příklad: *Reagovala už tak na připravované dělení státu, uvedl na tiskové konferenci předseda představenstva a. s. Fortuna Michal Horáček.* (mf920922_074#3)

	subjekt	subjekt v hlavní větě	frekvencované substantivum	nejčastější funktor	negace nejčastějšího funktoru	kolokace	vzdálenost	celkem
Fortuna _i	+1	+1	+1	+1			+1	5
kancelář _i	+1	+1	+1	+1			+2	6
Fortuna _j			+1		-1		+2	2
kancelář _j			+1		-1			0

Tab. 4.1: Hledání antecedentu pro anafor \emptyset_i

	subjekt	subjekt v hlavní větě	frekvencované substantivum	nejčastější funktor	negace nejčastějšího funktoru	kolokace	vzdálenost	celkem
kancelář _i	+1	+1	+1	+1			+1	5
Fortuna _j	+1		+1		-1		+1	2
\emptyset_i	+1			+1			+1	3
kancelář _j			+1		-1		+1	1
Fortuna _k			+1		-1		+2	2
konference					-1		+2	1

Tab. 4.2: Hledání antecedentu pro anafor \emptyset_j

Na tomto ilustračním příkladě je vidět koreferenční řetěz mezi *kancelář_i-svůj- \emptyset_i - \emptyset_j* . AČA ale pro něj nevytvořil řetěz. Předpokládá, že vyřeší každý druh anafory zvlášť a teprve na konci bude vytvořen koreferenční řetěz tam, kde odkazuje k jednomu antecedentů víc anaforů.

4.1.3 Výsledky obou řešení pro osobní zájmenou anaforu

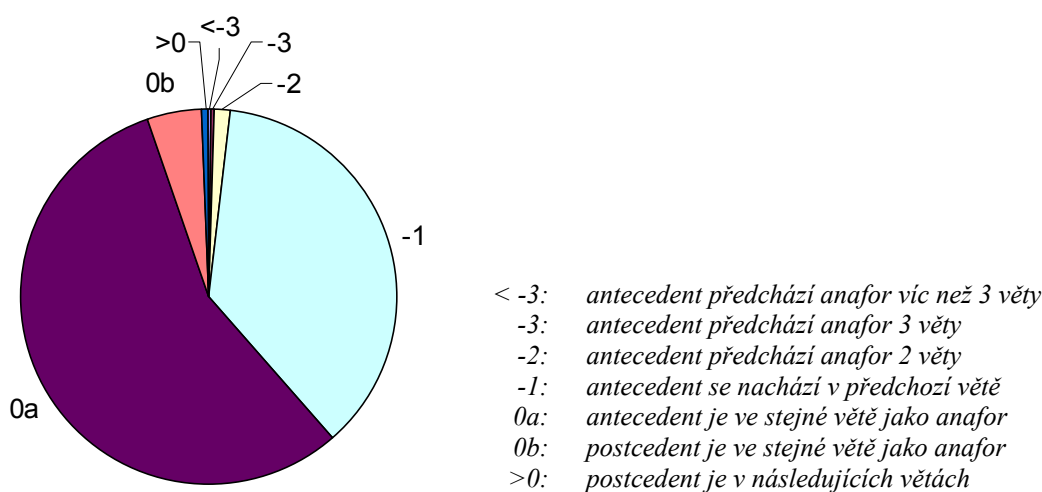
	verze 1	verze 2	verze 3	verze 4	verze 5

	pers	app	pers	app	pers	app	pers+app	pers	app	pers+app
precision	60,6%	62,1%	68,3%	66,2%	73,6%	55,3%	74,9%	76,2%	64,2%	68,4%
recall	51,2%	56,1%	51,2%	56,1%	75,8%	55,6%	74,9%	76,2%	64,2%	66,7%

pers anafory jsou osobní zájmena
app anafory jsou osobní posesivní zájmena

Tab. 4.3: Výsledky obou řešení pro osobní a posesivní zájmenou anaforu

Verze 1, 2 a 3 jsou různé experimenty s C4.5. Verze 1 vybírá kandidáty, které jsou pojmenovací sémantická substantiva, ze stejné věty jako anafor a ze tří předchozích vět. Verze 2 narozdíl od verze 1 už vybírá kandidáty jen ze stejné věty nebo z předchozí věty. Verze 3 bere všechna sémantická substantiva (kromě těch v 1. a 2. osobě) ze stejné věty nebo z předchozí věty jako kandidáty. Ve verzi 3 je také změněn seznam aktantů pro atribut *kandidát je aktant* a *anafor je aktant*: seznam aktantů se změnil na seznam nejčastějších funktorů u antecedentů osobní zájmenné anafory, který se skládá z ACT, PAT, ADDR a APP.



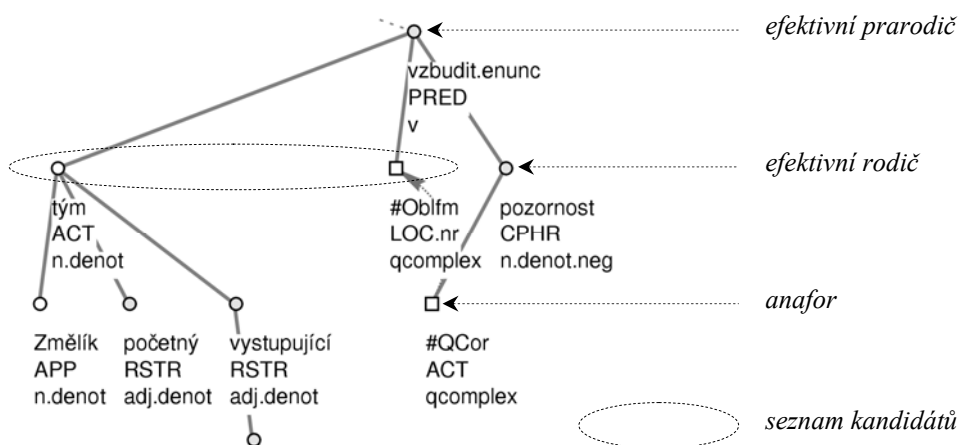
Obr. 4.6: Relativní četnost vzdálenosti antecedentu od osobního zájmenného anaforu v trénovacích datech PDT

Verze 4 je řešení pomocí ručně psaných pravidel. Verze 5 je řešení pomocí C4.5, které experimentuje s převodem ručně psaných pravidel do vektoru 7 atributů se dvěma třídami, protože nedosahuje nejlepšího výsledku, nebudeme ho dále rozebírat. Zkoušeli jsme vylepšit posesivní anaforu tím, že místo shody anaforu s kandidátem jsme vzali shodu rodiče anaforu s kandidátem (pro konstrukce typu: Potkali jsme *Jana* a *jeho* přítelkyni.), ale úspěšnost byla jen okolo 52%.

4.2 Nulová anafora kontroly

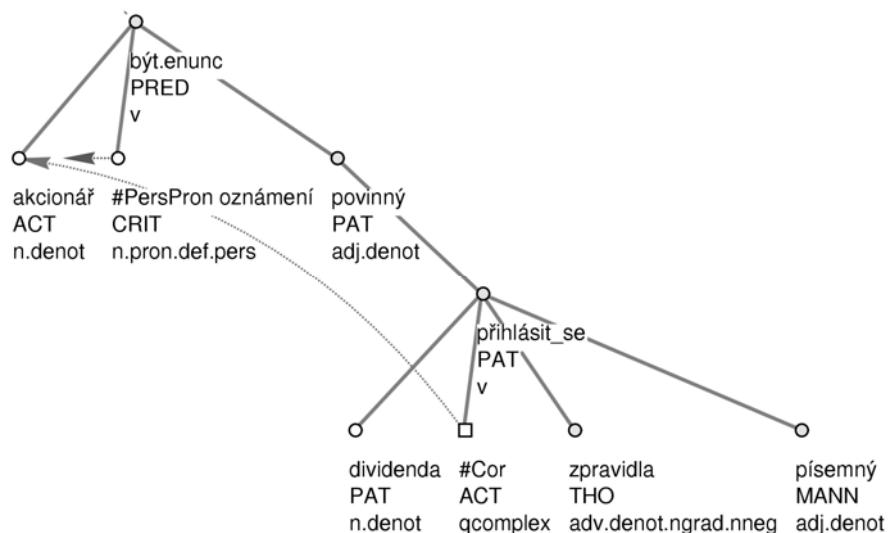
AČA sestaví seznam anaforů kontroly z uzlů s lematem *Cor* nebo *QCor*. Jelikož antecedentem kontroly je aktantem řídicího slovesa, najde AČA pro každý prvek ze seznamu jeho efektivního rodiče⁸ a efektivního prarodiče. (viz obr. 4.7) Je-li prarodič adjektivem, půjde o případ konstrukce "být odhodlaný/schopný/... udělat něco", proto půjde AČA ještě o jednu úroveň výš a dosadí za prarodiče praparodiče, za rodiče prarodiče. (viz obr. 4.8)

AČA pak vezme všechny efektivní potomky od prarodiče kromě rodiče anaforu jako seznam kandidátů. Je-li v seznamu kandidátů uzel s lematem *možný/nutný/třeba*, bude se antecedent nacházet právě mezi jeho potomky. Provede se výměna seznamu kandidátů za seznam potomků uzlu s lematem *možný/nutný/třeba*. (viz obr. 4.9)

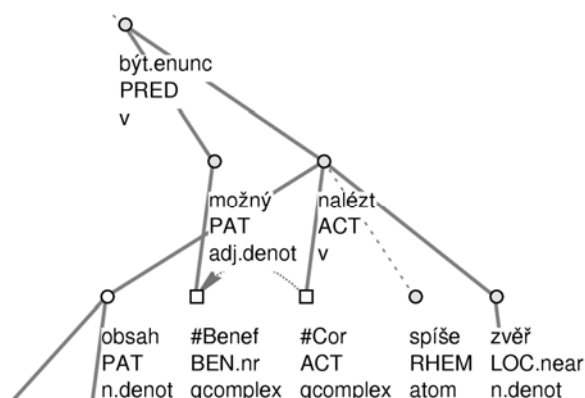


Obr. 4.7: Nulová anafora kontroly: *Změlíkův početný tým vystupující při nedávném desetiboji v Götzisu vzbudil pozornost.* (mf930709_075#8)

⁸ Efektivní rodič nějakého uzlu má s uzlem pravý závislostní vztah.



Obr. 4.8: Kontrola být + adjektivum: *Akcionář je podle něho povinen se o dividendy přihlásit zpravidla písemně.* (ln94204_160#13)



Obr. 4.9: Kontrola být + možný: *Vyšší obsahy jedovatých kovů je možné nalézt spíše u divoké zvěře.* (sample7#19)

Potom prochází AČA seznamem kandidátů a vytvoří potřebná data pro generování rozhodovacího stromu pomocí programu C4.5. Vektor nulové anafory kontroly se skládá z následujících atributů:

- *jedinečnost kandidáta:* Možné hodnoty {ano, ne}. Je-li v seznamu kandidátů jediný uzel, ten bude antecedentem; jinak ne.
- *kategorie prarodiče:* Možné hodnoty {V_ACT, V_ADDR, V_BEN, V_PAT, V_ORIG, V_ACT_ADDR, V_ACT_LOC, V_ACT_ORIG, V_ACT_PAT, V_ACT_BEN_PAT, V_ACT_ADDR_BEN, N_ACT, N_ADDR, N_BEN, N_PAT, b_mozny, b_adj, b_compl, other}. Řídící slovesa jsou rozdělena do skupin, v nichž je vztah mezi aktanty řídicího slovesa a anaforem kontroly stejný. O seznamech

sloves kontroly bude pojednáno níže. Hodnota *b_mozny* platí pro konstrukce typu „být možný/nutný/třeba...“; *b_adj* platí pro konstrukce „být odhodlaný/schopný/... udělat něco“; *b_compl* platí pro případy, kdy je efektivním rodičem doplněk (viz kapitola 4.7).

- *funktor kandidáta*: Možné hodnoty {ACT, AUTH, PAT, ADDR,...}
- *shoda funkтора kandidáta a funkтора anaforu s kategorií prarodiče*: Možné hodnoty {ano, ne}. Podmínky ke splnění shody jsou vysvětleny v následující podkapitole.
- *lema kandidáta*: {#Gen, #PersPron, #Benef, který, n, other}; n – sémantické substantivum.
- *anafora*: anafor odkazuje ke kandidátovi: ANO/NE

Kučová a kolektiv tvrdí: "Vztah kontroly je podmíněn lexikálním významem řídicího slovesa, předpokládáme tedy, že je potenciálně možné sestavit seznam sloves kontroly."⁹ AČA převzal seznam sloves kontroly z práce [Kučová et Žabokrtský, 2005], který byl získán ze slovníku VALLEX 1.0¹⁰. Původní seznam byl složen ze čtyř druhů sloves kontroly ADDR, PAT, BEN a ACT. Seznamy sloves jsou označeny podle funktoru jednoho ze svých aktantů, který bývá antecedentem nulové anafory kontroly.

Vysvětlíme to na slovesu *zakázat*, který je slovesem kontroly V_ADDR a má valenční rámec "zakázat něco; něco někde; někomu {něco; něco (u)dělat; aby}":

(4.1) Rodiče *mi* zakázali \emptyset odjet.

Nevyjádřený subjekt infinitivu *odjet* koreferuje k zájmenu *mi*, který je aktantem řídicího slovesa *zakázat* a má funktor ADDR.

Jelikož se řídicí sloveso může nominalizovat, sestavil AČA další čtyři seznamy substantiv nominalizovaných z existujících sloves (N_ACT, N_ADDR, N_BEN, N_PAT). Pomocí pozorování na testovacích souborech PDT, anotátorské příručky Anotace na tektogramatické rovině Pražského závislostního korpusu a Slovníku slovesných, substantivních a adjektivních vazeb a spojení přidali jsme další slovesa a substantiva do

⁹ Kučová, L. a kol.: Anotování koreference v Pražském závislostním korpusu, Universitas Carolina Pragensis, Praha 2003, str. 31

¹⁰ Valency Lexicon of Czech Verbs, Version 1.0 (VALLEX 1.0) je kolekce lingvisticky anotovaných dat a dokumentace, jejíž cílem je formální popis valenčních rámců českých sloves. Více na <http://ufal.mff.cuni.cz/~zabokrtsky/vallex/1.0/>

seznamů. Kromě toho byl seznam sloves kontroly rozšířen o další druhy, a to slovesa kontroly ORIG a slovesa kontroly, u nichž může být víc než jeden antecedent nebo antecedent může mít různé funkory: V_ACT_ADDR, V_ACT_LOC, V_ACT_ORIG, V_ACT_PAT, V_ACT_BEN_PAT, V_ACT_ADDR_BEN.

4.2.1 Podmínky shody funktora kandidáta a funktora anaforu s kategorií prarodiče

Kategorie b_mozny

Patří-li prarodič anaforu do kategorie b_mozny a funktor kandidáta je BEN, hodnota shody je pozitivní; jinak je negativní.

Kategorie b_adj

Prarodič je kategorie b_adj a funktor kandidáta je ACT, platí shoda; jinak ne.

Kategorie b_compl

Prarodič je b_compl a je-li doplněk vyjádřen infinitivem nebo určitým slovesem a kandidát má funktor PAT, pak je hodnota pozitivní; v ostatních případech, kdy doplněk je vyjádřen něčím jiným a funktor kandidáta je ACT, je hodnota také pozitivní; jinak je negativní.(viz kapitola 4.7.2)

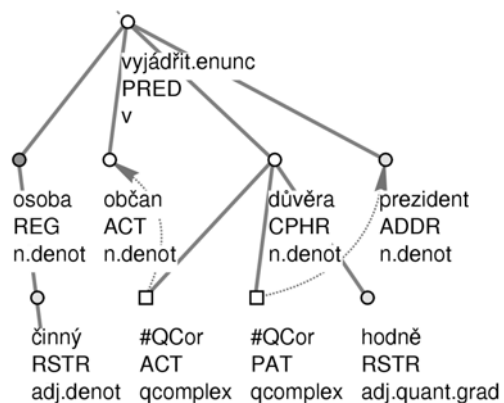
Kategorie V_ACT, V_ADDR, V_BEN, V_PAT, V_ORIG, N_ACT, N_ADDR, N_BEN, N_PAT

Patří-li prarodič do kategorie V_ACT, funktor kandidáta musí být ACT, aby splnila podmínku shody. Totéž platí pro ostatní kategorie: V_ADDR → ADDR, V_BEN → BEN, ..., N_ACT → ACT, N_BEN → BEN atd.

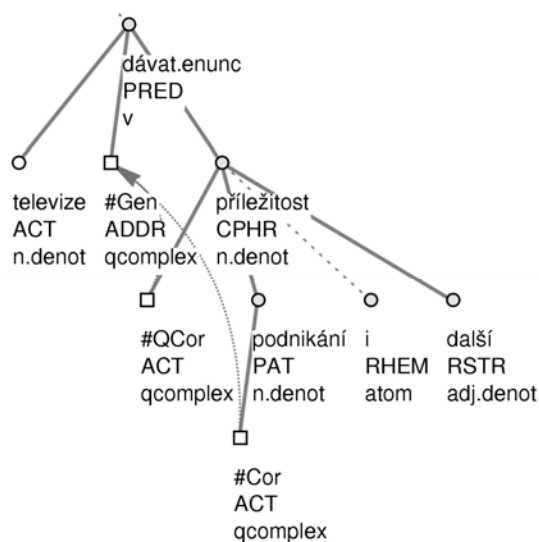
Kategorie V_ACT_ADDR

Anafor u slovesa kontroly V_ACT_ADDR může mít funktor ACT/ADDR/PAT (obr. 4.10 a 4.11). Anafor s funktorem ACT odkazuje k aktantu ACT řídicího slovesa, s

funktorem ADDR/PAT k aktantu ADDR (kromě případů *dát/dávat* + *naděje/šance/právo/...*¹¹, kdy anafor s funktorem ACT odkazuje k aktantu ADDR).



Obr. 4.10: Kontrola V_ACT_ADDR: *Z veřejně činných osob vyjádřili občané nejvíce důvěry prezidentovi.* (ln94203_43#17)

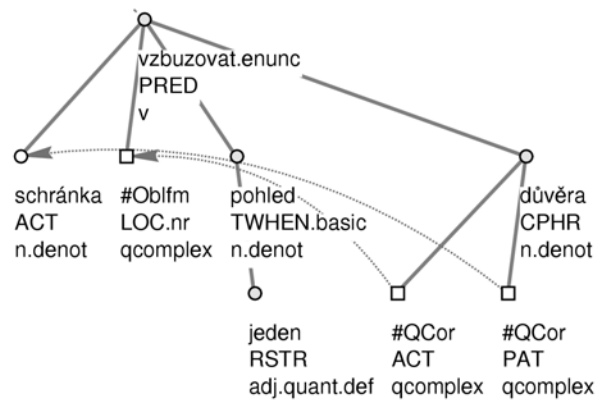


Obr. 4.11: Kontrola V_ACT_ADDR: *Televize dává i další příležitosti k podnikání.* (cmpr9415_047#18)

Kategorie V_ACT_LOC

Anafor s funktorem ACT u slovesa kontroly V_ACT_LOC odkazuje k aktantu LOC, anafor s funktorem PAT k aktantu ACT (obr. 4.12).

¹¹ Seznam substantiv byl získán pozorováním a je uveden v příloze jako n_ACT.



Obr. 4.12: Kontrola V_ACT_LOC: *Schránka vzbuzuje na první pohled důvěru.* (ln95047_145#1)

Kategorie V_ACT_ORIG

Má-li anafora konstrukce „dostat/získat doporučení/nabídka/odpověď...¹²“, musí platit: funktor kandidáta je ACT a funktor anaforu ADDR nebo funktor kandidáta je ORIG a funktor anaforu ACT, aby byla hodnota shoda pozitivní.

Má-li anafora konstrukce „dostat/získat pokuta¹³“, musí platit: funktor kandidáta je ACT a funktor anaforu PAT nebo funktor kandidáta je ORIG a funktor anaforu ACT, aby byla hodnota shoda pozitivní.

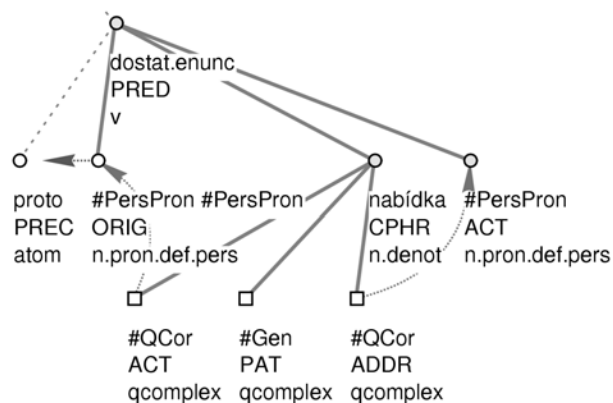
Má-li anafora konstrukce „dostat/získat návrh/souhlas¹⁴“, platí: funktor kandidáta je ORIG, funktor anaforu ACT.

V ostatních případech, kde je funktor kandidáta ACT, hodnota je pozitivní, jinak ne.

¹² Seznam substantiv je uveden v příloze jako n_ADDR.

¹³ Seznam je uveden v příloze jako n_PAT.

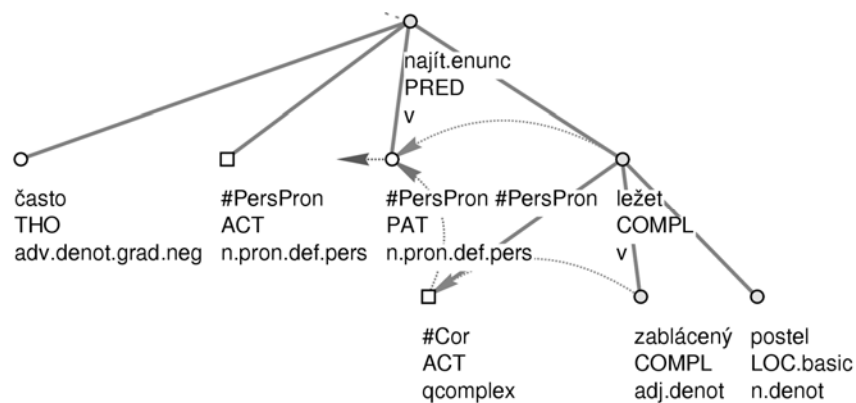
¹⁴ Seznam je uveden v příloze jako n_ORIG.



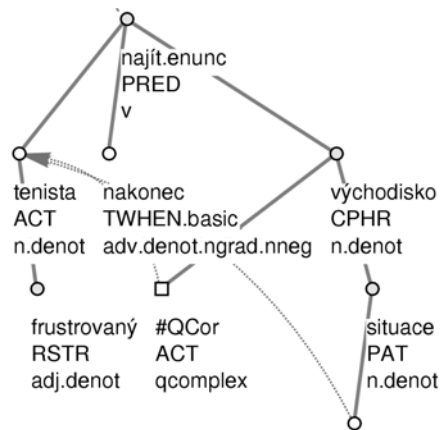
Obr. 4.13: Kontrola V_ACT_ORIG: *Proto jsme od nich dostali nabídku my.* (mf930709_100#6)

Kategorie V_ACT_PAT

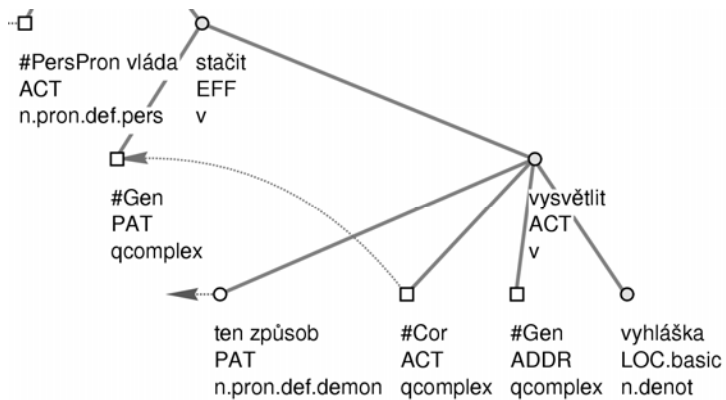
Antecedentem u slovesa kontroly V_ACT_PAT *najít* + *sloveso* je aktant PAT, u konstrukce *najít* + *substantivum* je antecedentem aktant ACT (obr. 4.14 a 4.15). U slovesa kontroly V_ACT_PAT *stačit* rozlišujeme dva případy (viz obr. 4.16 a 4.17), kdy je ve větě vyjádřený (anafor odkazuje k aktantu ACT) nebo všeobecný/nevyjádřený subjekt (anafor odkazuje k aktantu PAT).



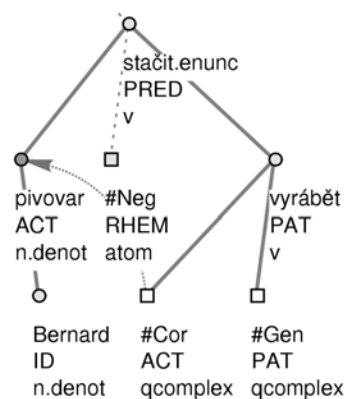
Obr. 4.14: Kontrola V_ACT_PAT: *Často jsme ho našli, jak leží zablácený v posteli.* (lnd94103_087#86)



Obr. 4.15: Kontrola V_ACT_PAT: *Frustrovaný tenista nakonec našel východisko ze své situace:* (In94204_129#15)



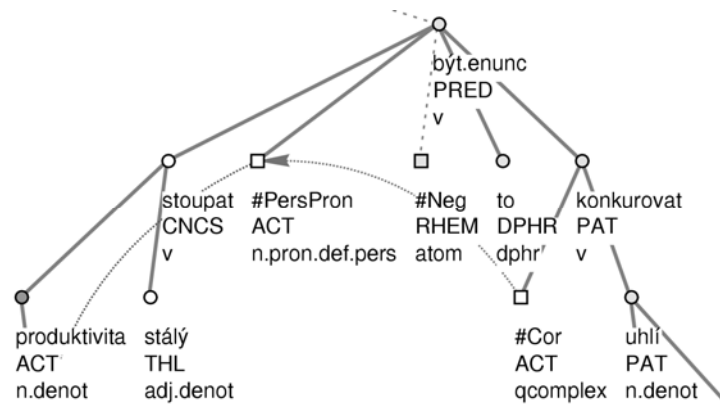
Obr. 4.16: Kontrola V_ACT_PAT: *"Domnívá se, že to stačí vysvětlit ve vyhlášce," sdělil Morávek.* (In94205_142#9)



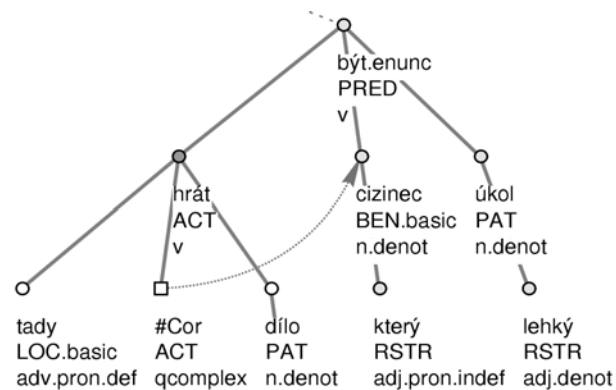
Obr. 4.17: Kontrola V_ACT_PAT: *Pivovar Bernard nestačí vyrábět.* (In94204_27#1)

Kategorie V_ACT_BEN_PAT

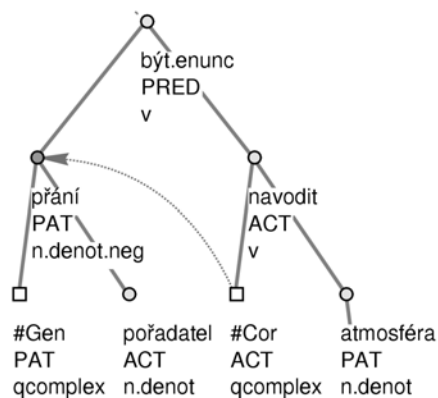
U slovesa kontroly V_ACT_BEN_PAT *být* zase rozlišujeme, kdy má efektivní rodič funktor ACT nebo jiný (obr. 4.18, 4.19 a 4.20).



Obr. 4.18: Kontrola V_ACT_BEN_PAT: *Ačkoli produktivita evropských dolů stále stoupá, nejsou s to konkurovat levnému uhlí ze zámoří.* (ln94204_19#8)



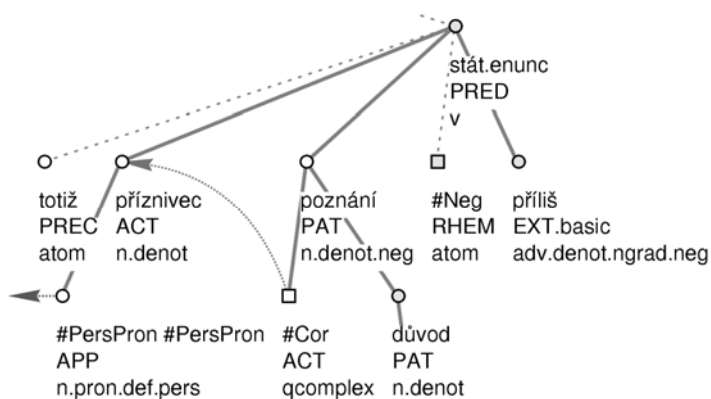
Obr. 4.19: Kontrola V_ACT_BEN_PAT: *Hrát zde tak důvěrně známé dílo je pro každého cizince nelehký úkol.* (ln94209_75#7)



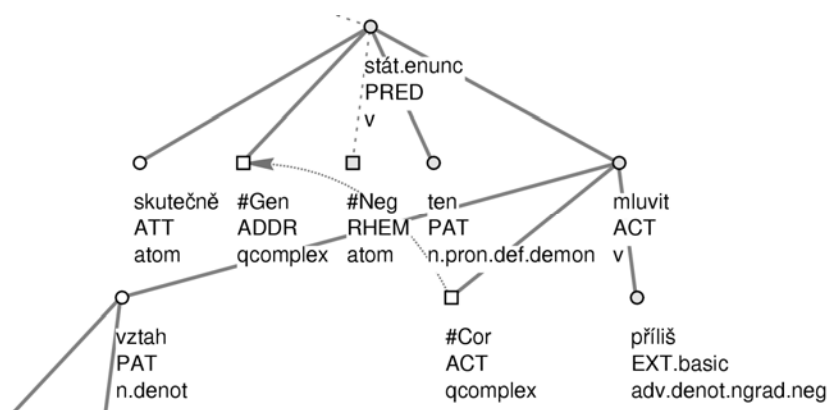
Obr. 4.20: Kontrola V_ACT_BEN_PAT: *Přáním pořadatelů je navodit atmosféru amerického Woodstocku.* (In94204_148#4)

Kategorie ACT – ADDR – BEN

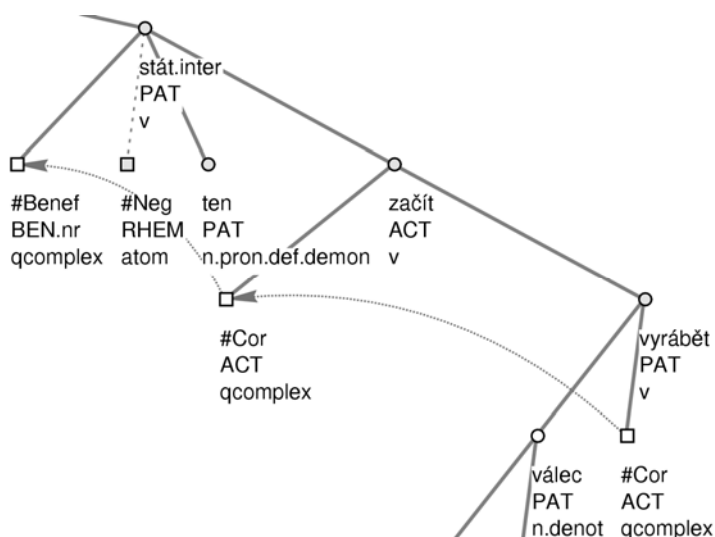
U slovesa kontroly V_ACT_ADDR_BEN předpokládáme, že v seznamu kandidátů je právě jen jeden z těchto tří aktantů (obr. 4.21, 4.22 a 4.23).



Obr. 4.21: Kontrola V_ACT_ADDR_BEN: *Jejich příznivci totiž o poznání skutečných důvodů příliš nestojí.* (In94204_139#48)



Obr. 4.22: Kontrola V_ACT_ADDR_BEN: *O našich vztazích skutečně nestojí za to příliš mluvit.* (ln94207_65#38)



Obr. 4.23: Kontrola V_ACT_ADDR_BEN: *Proto se táži všech majitelů, či ředitelů strojírenských firem a kovovýroby: "Nestálo by za to, začít lázeňské válce na tuhá paliva vyrábět?"* (cmpr9410_008#24)

4.2.2 Výsledek AČA pro nulovou anaforu kontroly

AČA uspěl s precision = 96,2%; recall = 87,3% pro nulovou anaforu kontroly

4.3 Reflexivní zájmenná anafora

Nad uzlem reprezentující reflexivní zájmena *se, si, svůj*¹⁵ najde AČA nejbližší určité sloveso nebo uzel s funktoem DENOM¹⁶. Mezi efektivními potomky určitého slovesa

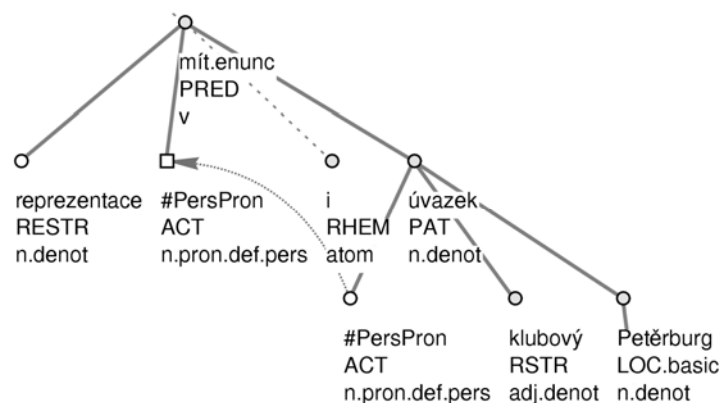
¹⁵ To jsou uzly s lematem #PersPron a gramatémem *person* s hodnotou *inher*

¹⁶ Funktor DENOM (denomination) je funktoer pro efektivní kořen nezávislé nominativní klauze, která není vsuvkou ([PDT-manuál])

nebo uzlu DENOM se nachází antecedent. Před vytvořením vstupních dat pro C4.5 prochází AČA seznam kandidátů tvořený z efektivních potomků slovesa, kteří jsou sémantická substantiva¹⁷, a určí subjekt věty a počet osobních zájmen.

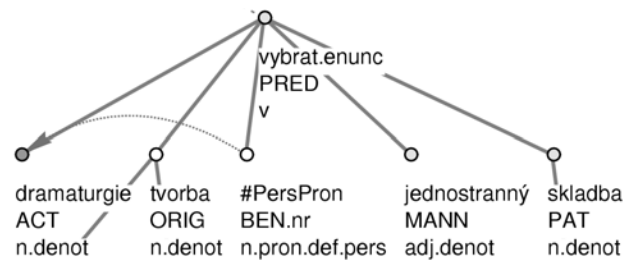
AČA používá následující atributy pro dvojici reflexivního zájmena a kandidáta na antecedent:

- *jediné osobní zájmeno*: Možné hodnoty {ano, ne}. Je-li kandidát jediným osobním zájmenem v seznamu kandidátů, pak je hodnota pozitivní, jinak je negativní.
- *kandidát je subjekt*: Možné hodnoty {ano, ne}. Je-li kandidát subjektem věty nebo věta nemá subject a kandidát má funktor ACT, pak je hodnota pozitivní, jinak je negativní.
- *kandidátův funktor*: Možné hodnoty {ACT, AUTH, PAT,...}.
- *kandidát je aktant*: Možné hodnoty {ano, ne}. Je-li kandidát aktantem řídicího slovesa, hodnota je pozitivní, jinak je negativní.
- *kategorie anaforu*: Možné hodnoty {se-si, svůj}.



Obr. 4.24: Reflexiva: *Kromě reprezentace mám i svůj klubový úvazek v Sankt Petěrburgu.* (ln94207_65#34)

¹⁷ Uzel, který má atribut *sempos* s hodnotou začínající na *n*, je sémantické substantivum ([PDT-manuál])



Obr. 4.25: Reflexiva: *Dramaturgie si z půlstoleté Stahuljakovy tvorby jednostranně vybrala skladby z dvacátých let.* (ln95047_051#7)

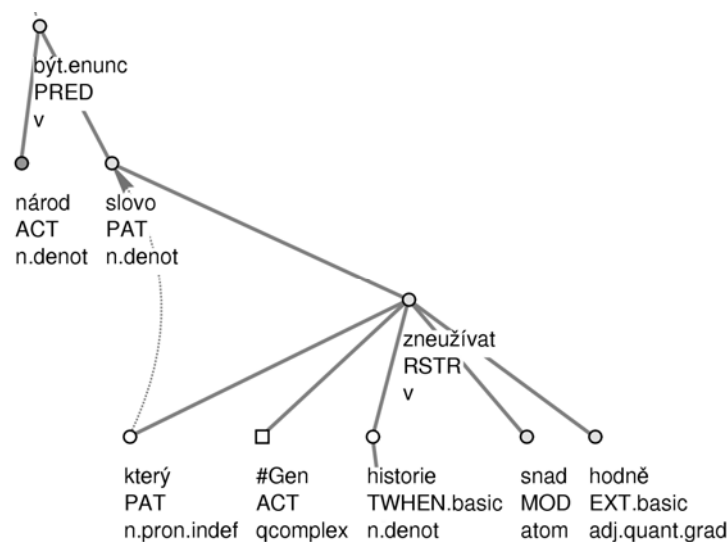
4.3.1 Výsledek pro reflexivní zájmenou anaforu

AČA uspěl s precision = 97,9%; recall = 96,2% pro anafory *se/si* a precision = 98%; recall = 96,3% pro anafory *svůj*.

4.4 Relativní zájmenná anafora

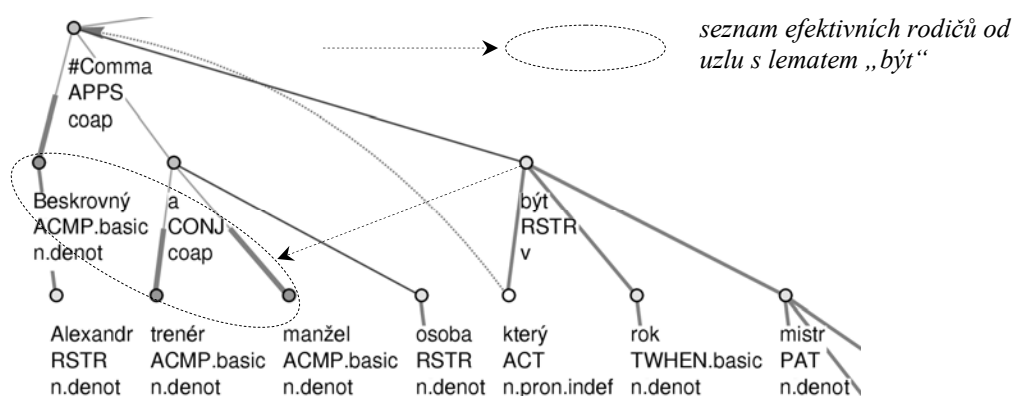
AČA rozdělila relativní zájmenou anaforu na dva dílčí problémy. Ten první je hledání antecedentů pro relativní zájmena kromě výrazu *což*. Ten druhý je hledání antecedentů právě pro spojovací výraz *což*. Obě řešení jsou založena na ručně psaných pravidlech.

4.4.1 Řešení pro relativa kromě *což*



Obr. 4.26: Relativa: *Národ je slovo, jež bylo v lidské historii snad nejvíc zneužíváno...* (ln94207_92#17)

Jako anafory AČA vyhledá všechny uzly reprezentující vztažná zájmena, vztažná zájmenná adverbia a vztažné zájmenné číslovky¹⁸ kromě těch, které jsou frazémem, etickým dativem, „falešným podmětem“ nebo součástí ustáleného spojení¹⁹, a kromě těch, které mají na analytické úrovni lema *což*. Nad každým anaforem najde nejbližší určité sloveso. Má-li sloveso jediného efektivního rodiče, ten bude antecedentem. Je-li efektivních rodičů víc než jeden, jedná se o případ, kdy anafor odkazuje k podstromu obsahujícímu všechny antecedenty (viz obr. 4.20). V tom případě musí AČA najít nad určitým slovesem nejbližší uzel, v jehož podstromu jsou efektivní rodiče slovesa. AČA nepoužívá C4.5 k určení relativní zájmenné anafory díky struktuře PDT, kde je vztah antecedentu a vztažného anaforu „téměř“ jednoznačně označen.



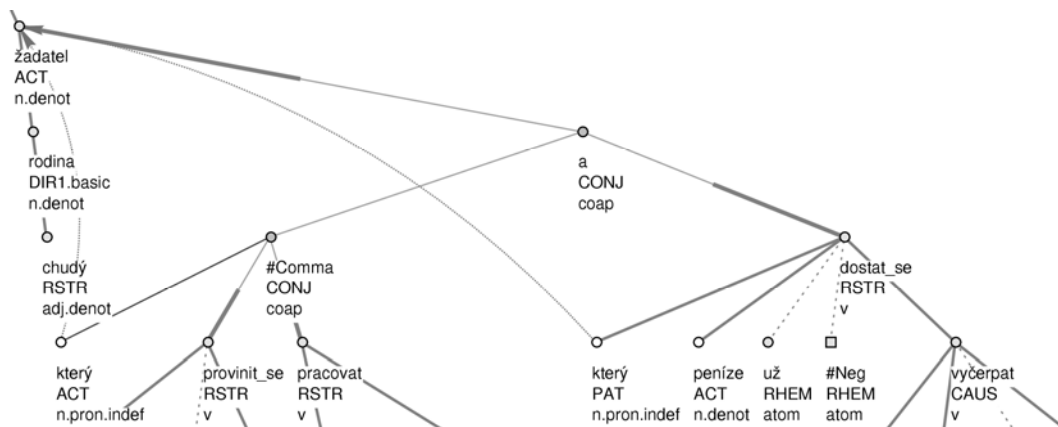
Obr. 4.27: Relativa: *S Alexandrem Beskrovným, trenérem a manželem v jedné osobě, který byl v roce 1978 juniorským mistrem Evropy v trojskoku, máme v Bratislavě byt a USK se se mnou dohodl na hostování do konce roku.* (ln94208_116#8)

Před tím, než jsme dospěli k tomuto řešení relativní zájmenné anafory, jsme zkoušeli i pravidlo *RelativeClauseRule* od [Kučová], které říká: „Najdi nejbližší uzel s funktorem RSTR (kořen vztažné věty); substantivum nebo zájmeno nad ním vyber jako antecedent.“²⁰ Tento způsob ale nepočítá s případy, kdy je antecedentem celý podstrom nebo kdy je kořenem vztažné věty spojka nebo čárka (obr. 4.28).

¹⁸ To jsou uzly, které mají gramatém *indeftype* s hodnotou *relat*

¹⁹ To jsou uzly, které mají funktor ETHD, DPHR nebo INTF. (PDT-manuál)

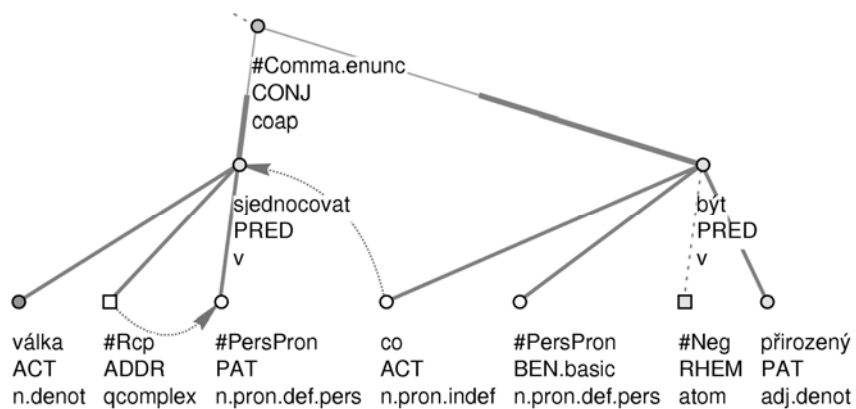
²⁰ Kučová, L. a kol.: Anotování koreference v Pražském závislostním korpusu, Universitas Carolina Pragensis, Praha 2003, str. 18



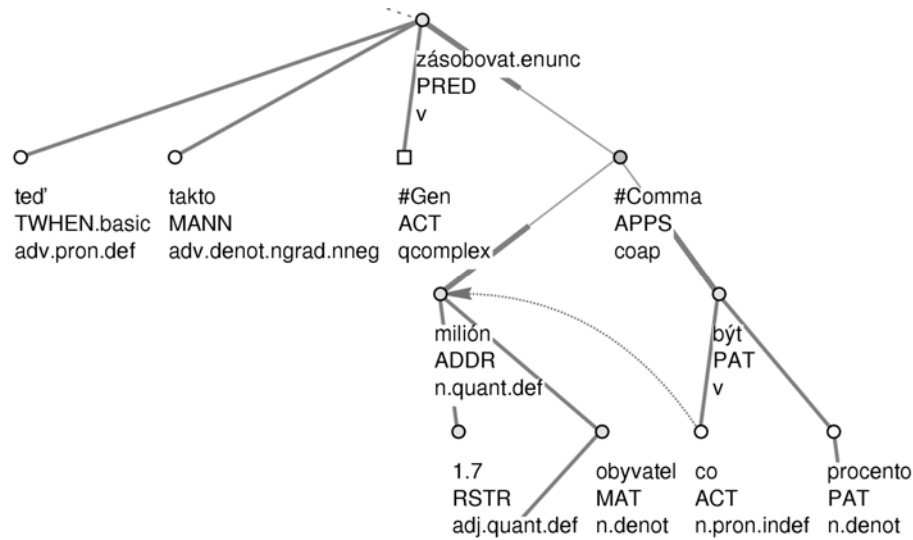
Obr. 4.28: Relativa: *Na tom by nebylo nic špatného, kdyby nebylo žadatelů z chudých rodin, kteří se nikdy neprovinili proti zákonu, pracují sedm dní v týdnu za minimální mzdu, a na které se peněz už nedostane, protože roční podíl je vyčerpán na ty ještě "chudší", kteří za státní útraty nedělají nic.* (ln94103_089#19)

4.4.2 Řešení pro což

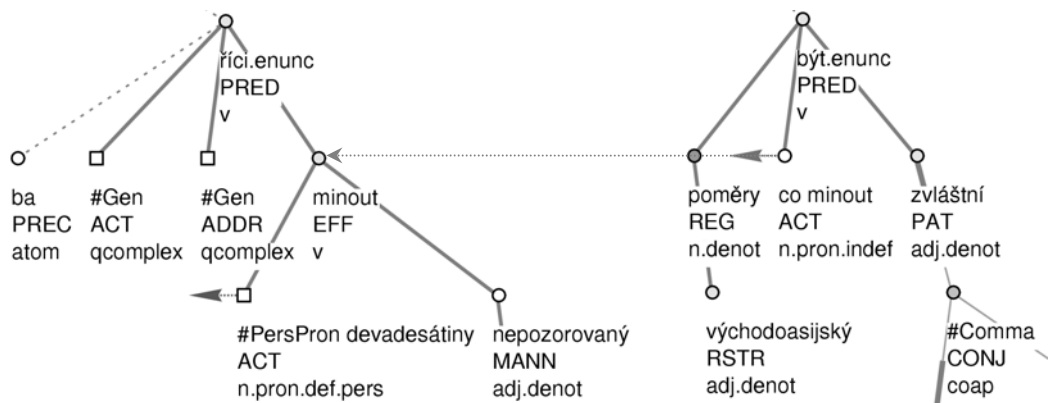
Je-li prarodič anaforu *což* čárka nebo pomlčka, určí AČA levého bratra od rodiče *což* za antecedent (obr. 4.29 a 4.30). Jinak bude nejbližší určité sloveso předcházející rodiče *což* určen jako antecedent (obr. 4.31).



Obr. 4.29: Což: *Válka nás sjednocuje, což pro nás není přirozené.* (ln94207_92#43)



Obr. 4.30: Což: Nyní je takto zásobováno 1.7 milionu obyvatel ČR, což je téměř 17 procent. (ln94202_84#3)



Obr. 4.31: Což: Ba dá se říci, že minuly takřka nepozorovaně. Což je na východoasijské poměry velmi, velmi zvláštní. (ln94200_123#7-8)

4.4.3 Výsledek pro relativní zájmennou anaforu

AČA uspěl s precision = recall = 99,6% pro relativní zájmennou anaforu kromě *což*. Výsledek pro *což* je: precision = recall = 96,6%.

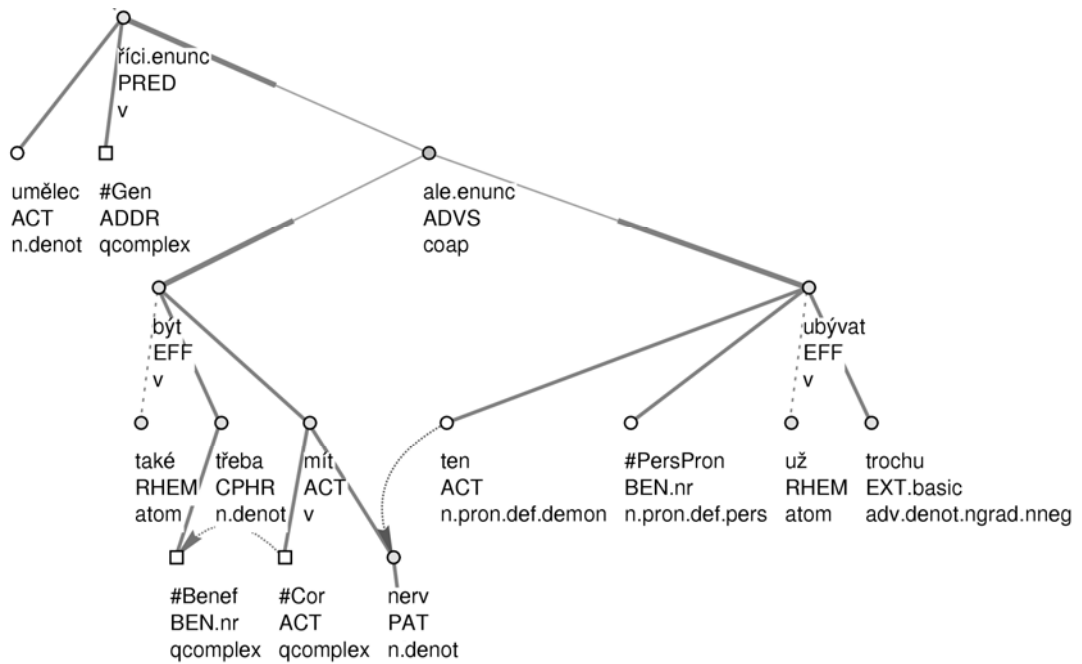
4.5 Demonstrativní zájmenná anafora

AČA hledá antecedenty pro všechna odkazující ukazovací zájmena kromě zájmena *to*, které obvykle odkazuje k předchozí větě nebo k předcházejícím větám.

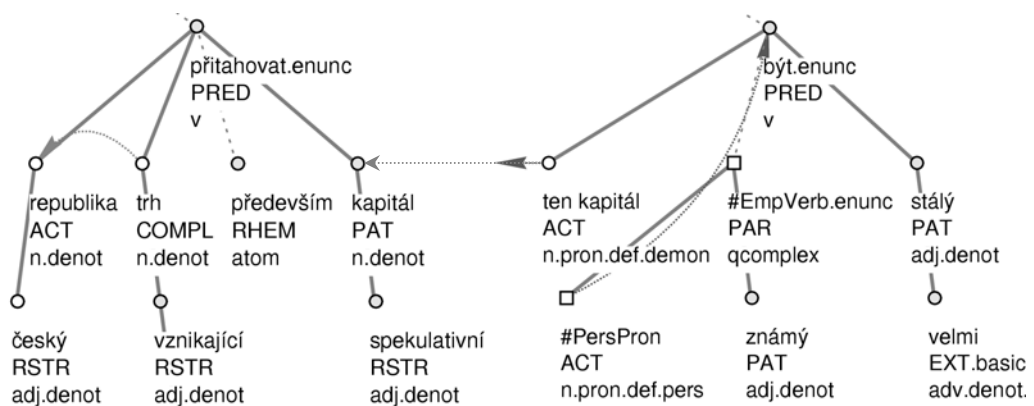
Pravidlo pro demonstrativní zájmennou anaforu kromě zájmena *to* je takové: Najdi nejbližší sémantické substantivum předcházející ukazovací zájmeno, které se s ním

shoduje v čísle a rodě (obr. 4.32 a obr. 4.33). V případě, že je ukazovací zájmeno rozvíjeno adjektivem a je v singuláru, může antecedent být v plurálu (obr. 4.34).

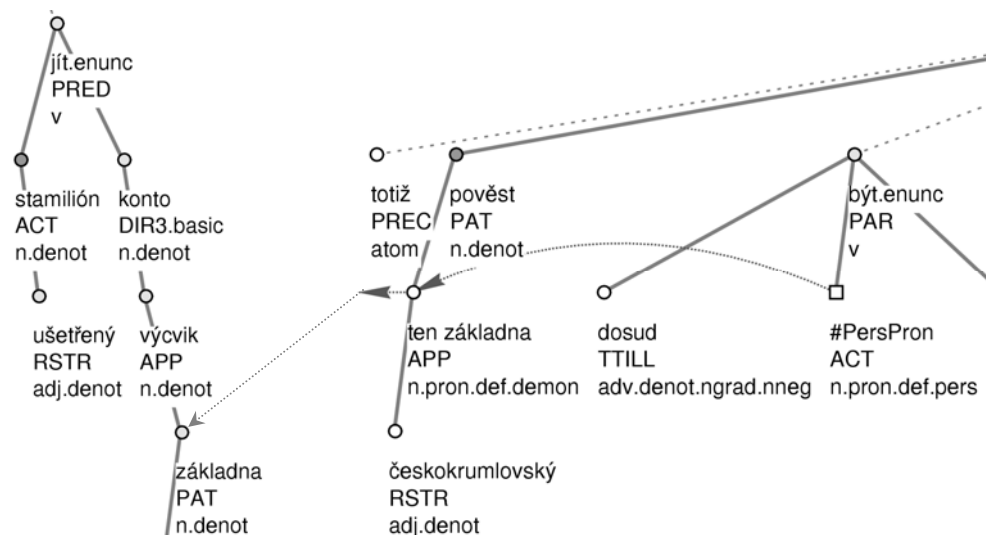
Toto pravidlo samozřejmě neurčí správně antecedenty pro případy, kdy je antecedent celá klauze nebo celá věta. Problém hledání takových antecedentů přesahuje rámec naší práce, protože k vyřešení vyžaduje znalosti kontextu a světa, které nemáme k dispozici.



Obr. 4.32: Demonstrativa: *Také je třeba mít zdravé nervy, ale těch mi už trochu ubývá, řekl umělec.* (ln94209_76#20)



Obr. 4.33: Demonstrativa: *Česká republika jako nově vznikající trh přitahuje především spekulativní kapitál. Ten, jak známo, může být velmi nestálý.* (ln94205_139#65)

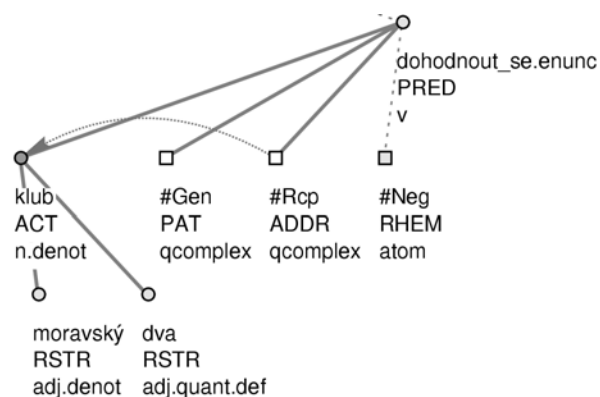


Obr. 4.34: Demonstrativa: *Ušetřené stamilióny půjdou na konto výcviku armádních základen. Pověst té českokrumlovské (dosud je v podřízenosti sekce zahraničních vztahů ministerstva obrany) byla totiž zřejmě vystavěna tak trochu na účet živořících tankistů, pěšáků či dělostřelců.* (ln94207_36#40-41)

4.5.1 Výsledek pro demonstrativní zájmenou anaforu

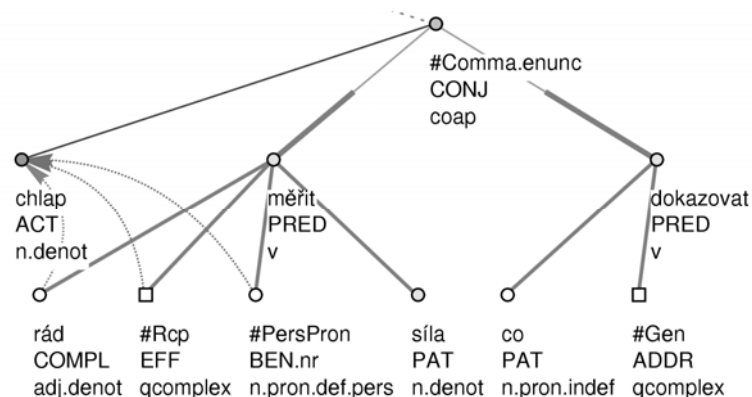
AČA uspěl s precision = recall = 72,6%.

4.6 Nulová anafora reciprocit

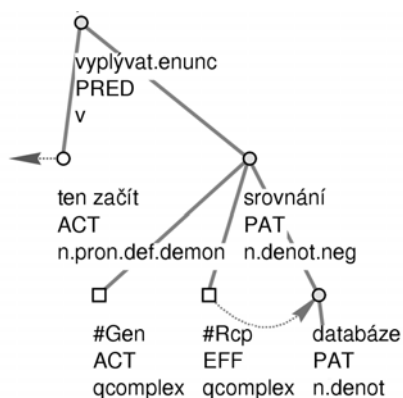


Obr. 4.35: Reciprocita: *Dva moravské kluby se nedohodly* (mf930709_003#1)

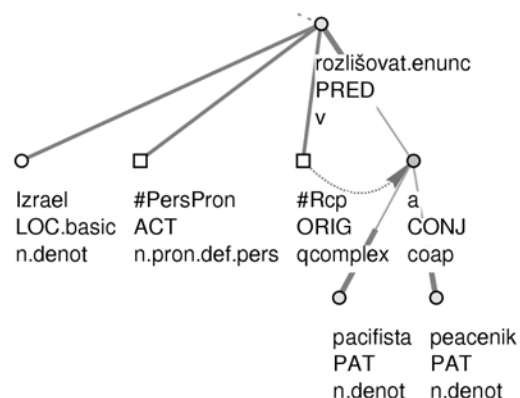
Má-li rodič tektogramatické lema koordinovat/měřit/srovnat/...²³ a funktor anaforu je EFF, určí AČA kandidát s funktoem ACT jako antecedent (obr. 4.38); jinak odkazuje anafor s funktoem EFF/ORIG k sourozenci s funktoem PAT (obr. 4.39 a 4.40).



Obr. 4.38: Reciprocita: *Chlapi si rádi měří síly, něco dokazují.* (ln94103_087#129)



Obr. 4.39: Reciprocita: *Vyplývá to ze srovnání obou databází.* (ln95045_101#7)



Obr. 4.40: Reciprocita: *V Izraeli rozlišujeme mezi pacifisty a peaceniky.* (ln94207_92#11)

²³ Seznam je uveden v příloze jako R_ACT_EFF.

Nepatří-li rodič do žádné ze dvou uvedených skupin tektogramatických lemat, bude kandidát s funktorem ACT vybrán jako antecedent.

4.6.1 Výsledek pro nulovou anaforu reciprocit

AČA uspěl s precision = recall = 94,7%.

4.7 Doplněk

Doplněk je rozvíjející větný člen, který závisí na slovese (zpravidla v přísudku), ale zároveň se vztahuje ke jménu (nejčastěji v podmětu nebo v předmětu)²⁴. V PDT je první závislost znázorněna hranou závislostního stromu, druhá závislost je znázorněna pomocí šipky podobným způsobem jako zachycení koreferenčního vztahu.

AČA navrhuje řešení pro dva různé druhy doplňku:

- doplněk vyjádřený substantivem, adjektivem, zájmenem nebo číslovkou²⁵
- doplněk vyjádřený slovesem nebo lematem *rád, sám*

4.7.1 Doplněk vyjádřený substantivem, adjektivem, zájmenem nebo číslovkou

AČA zvolil za kandidáty na „sémanticky řídicí“ substantivum sourozence uzlu s doplňkem, které jsou sémantickými substantivy. Pro každou dvojici kandidát – doplněk používá atributy:

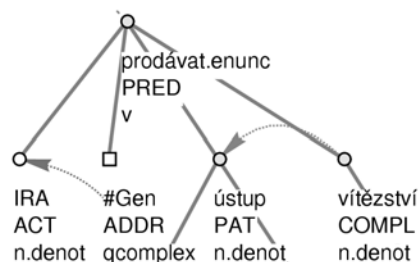
- *jediný kandidát*: Možné hodnoty {ano, ne}.
- *rod kandidáta*: Možné hodnoty {M, I, F, N}. M – mužský neživotný; I – mužský životný; F – ženský; N – střední rod.
- *rod doplňku*: Možné hodnoty {M, I, F, N}.
- *číslo kandidáta*: Možné hodnoty {S, P, D}. S – singulár; P – plural; D – duál.
- *číslo doplňku*: Možné hodnoty {S, P, D}.
- *pád kandidáta*: Možné hodnoty {1, 2, 3, 4, 5, 6, 7}.
- *pád doplňku*: Možné hodnoty {1, 2, 3, 4, 5, 6, 7}.

²⁴ Havránek, B – Jedlička, A.: Stručná mluvnice česká, Fortuna, Praha 1998, str. 165

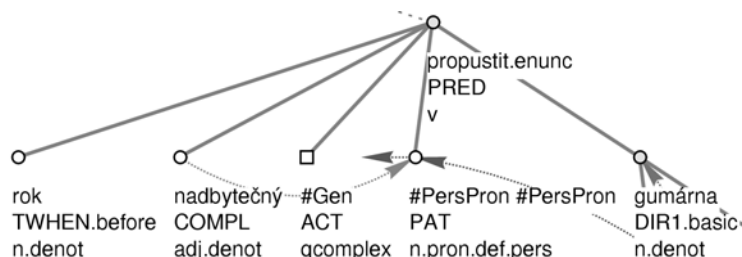
²⁵ To jsou uzly, které mají gramatém s hodnotou *n*, *adj.denot* nebo *adj.quant* ([PDT-manuál])

- *shoda v rodě*: Možné hodnoty {ano, ne}.
- *shoda v čísle*: Možné hodnoty {ano, ne}.
- *shoda v pádě*: Možné hodnoty {ano, ne}.
- *shoda v předložce*: Možné hodnoty {ano, ne}. Pro případ kdy je ve větě víc substantiv se stejným pádem jako doplněk.
- *doplněk se spojkou*: Možné hodnoty {ano, ne}. Je-li doplněk se spojkou *jako* (*jakožto*, *coby*), pak je hodnota pozitivní; jinak je negativní.
- *slovní druh doplňku*: Možné hodnoty {n, adj, pron, num}; n – substantivum; adj – adjektivum; pron – zájmeno; num – číslovka.

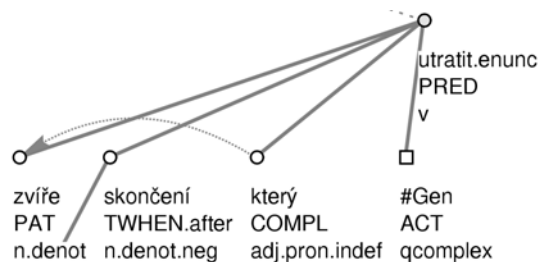
Doplňky vyjádřené adjektivem, zájmenem nebo číslovkou se shodují se svým substantivem v rodě (a životnosti), čísle i v pádě. U doplňků vyjádřených substantivem platí toto: je-li doplněk se spojkou *jako* (*jakožto*, *coby*), shoduje se s „řídícím“ substantivem v pádě; jinak je pád doplňku dán vazbou slovesa. Předpokládáme, že kdyby se podařilo sestavit seznam sloves pro doplněk, plnící podobnou funkci jako seznam sloves kontroly, určilo by se „řídící“ substantivum lépe.



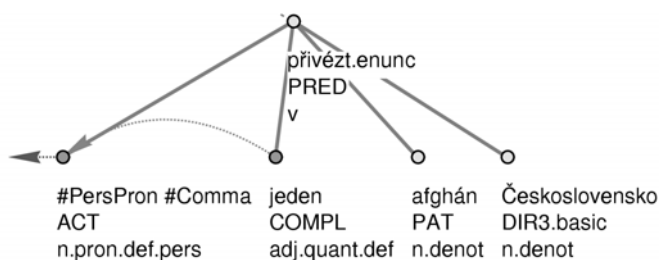
Obr. 4.41: Doplněk vyjádřený substantivem: *IRA prodává svůj ústup z ozbrojených pozic jako konečné vítězství.* (ln94209_50#14)



Obr. 4.42: Doplněk vyjádřený adjektivem: *Před rokem jej jako nadbytečného propustili z púchovských gumáren, kde pracoval dvacet let.* (mf930709_113#11)



Obr. 4.43: Doplněk vyjádřený zájmenem: *Zvířata jsou po skončení pokusu všechna utracena.* (In94200_124#21)



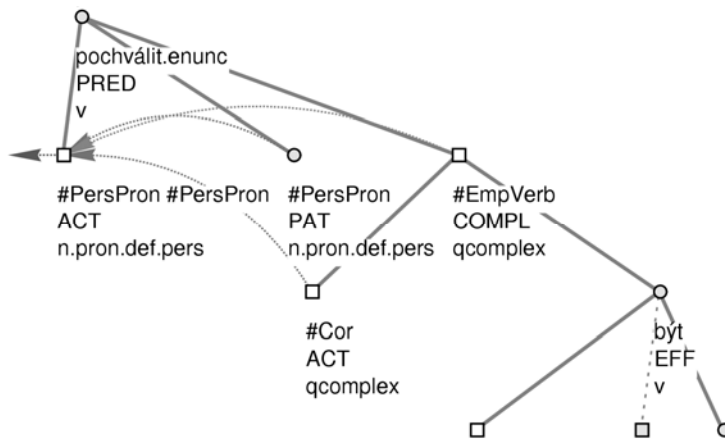
Obr. 4.44: Doplněk vyjádřený číslovkou: *On první přivezl afghány do Československa.* (In94103_087#79)

4.7.2 Doplněk vyjádřený slovesem nebo lematem *rád, sám*

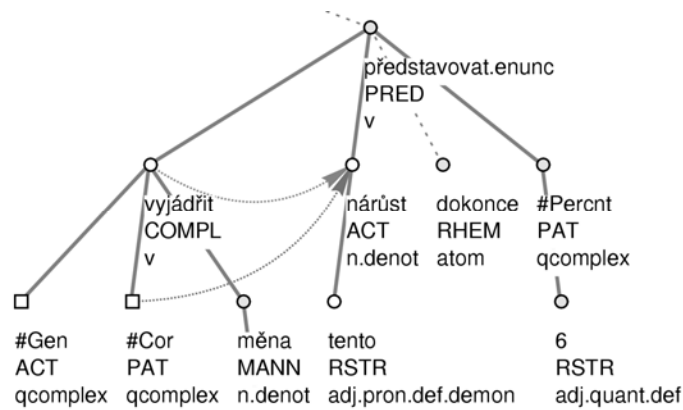
Seznam kandidátů na „řídící“ substantivum u doplňku vyjádřeného slovesem nebo lematem *rád, sám* tvoří sourozence uzlu doplňku, které jsou sémantikými substantivy a mají funktor ACT nebo PAT. AČA definuje pro dvojici kandidát – doplněk následující atributy:

- *kategorie doplňku:* Možné hodnoty {rad, sam, emp, trans, inf, part, fin}; rad – doplněk vyjádřený lematem *rád, sám* – *sám*; emp – doplněk s lematem *#EmpVerb*²⁶; trans – doplněk vyjádřený přechodníkem, inf – infinitivem, part – participiem; fin – doplněk vyjádřený určitým slovesem (vedlejší větou doplňkovou).
- *funktor kandidáta:* Možné hodnoty {ACT, PAT}.
- *shoda kategorie doplňku s funktořem kandidáta:* Možné hodnoty {ano, ne}. Je-li doplněk vyjádřen infinitivem nebo určitým slovesem a kandidát má funktor PAT, pak je hodnota pozitivní; v ostatních případech, kdy doplněk je vyjádřen něčím jiným a funktor kandidáta je ACT, je hodnota také pozitivní; jinak je negativní.

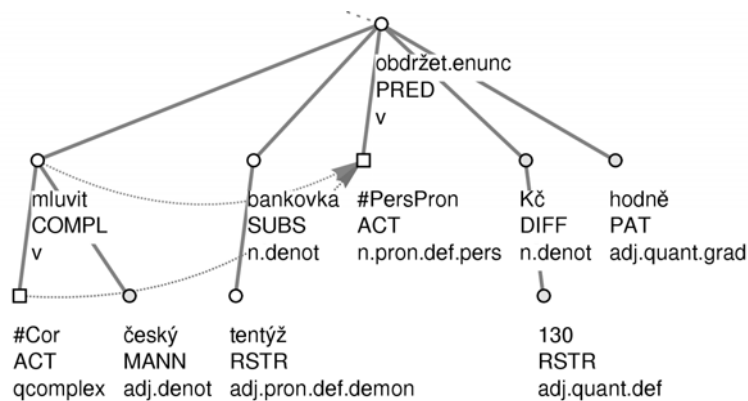
²⁶ Uzel s lematem *#EmpVerb* je elipsa řídícího slovesa ([PDT-manuál])



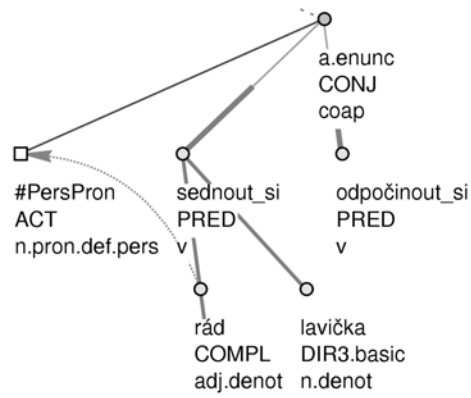
Obr. 4.45: Doplněk s lematem #EmpVerb: *Nejsem udavač, pochválil se.* (ln94204_34#15)



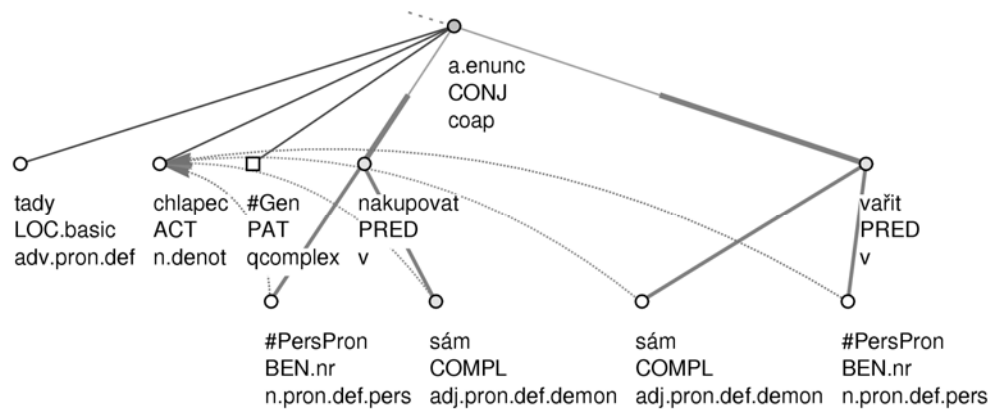
Obr. 4.46: Doplněk vyjádřený participiem: *Vyjádřen v místních měnách, tento nárůst představuje dokonce 6 %.* (ln94204_16#4)



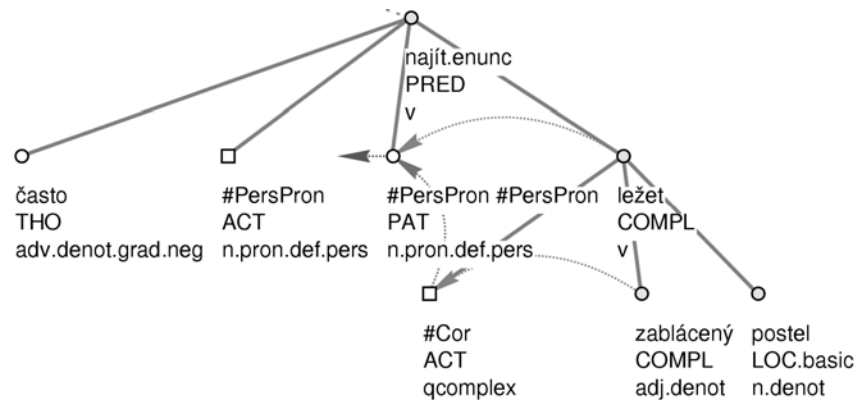
Obr. 4.47: Doplněk vyjádřený přechodníkem: *Mluvě česky, za tutíž bankovku obdržím o 130 Kč více.* (ln94205_136#5)



Obr. 4.48: Doplněk vyjádřený lematem rád: *Docela rád si sednu na lavičku a odpočinu si.* (ln94208_129#20)



Obr. 4.49: Doplněk vyjádřený lematem sám: *Tady si chlapci sami nakupují a sami si vaří.* (ln95047_061#88)



Obr. 4.50: Vedlejší věta doplňková: *Často jsme ho našli, jak leží zablácený v posteli.* (ln94103_087#86)

4.7.3 Výsledek pro doplňky

AČA uspěl s precision = 87,2%; recall = 85% pro doplňky vyjádřené sémantickým substantivem a precision = 100%; recall = 92,5% pro doplňky vyjádřené slovesem, *rád*, *sám*.

Kapitola 5

Závěr

	precision	recall
osobní zájmenná anafora	75,8%	75,8%
posesivní zájmenná anafora	64,1%	64,1%
nulová anafora kontroly	96,2%	87,3%
reflexivní zájmenná anafora	98%	96,2%
relativní zájmenná anafora kromě <i>což</i>	99,6%	99,6%
relativní zájmenná anafora <i>což</i>	96,6%	96,6%
demonstrativní zájmenná anafora kromě <i>to</i>	72,6%	72,6%
nulová anafora reciprocity	94,7%	94,7%
doplňek vyjádřený sémantickým substantivem	87,2%	85%
doplňek vyjádřený slovesem, <i>rád, sám</i>	100%	92,5%

Tab. 5.1: Vyhodnocení úspěšnosti AČA

Cílem naší diplomové práce byl návrh souboru pravidel pro analýzu anafor v českém jazyce. Vzhledem k používaným datům s ručně anotovanými koreferencemi v PDT jsme omezili na analýzu anafory osobní, posesivní, demonstrativní, reflexivní, relativní (včetně spojovacího výrazu *což*), nulové anafory kontroly a reciprocity. Dodatečně jsme si určili cíl analyzovat závislostní vztah mezi doplňky a jmény.

Pro osobní a posesivní jsme použili jak rozhodovací stromy C4.5, tak jsme aplikovali i ručně psaná pravidla. Navržené atributy pro rozhodování a ručně psaná pravidla byly inspirovány částečně Lappin & Leassovým, Mitkovovým algoritmem a teorií o aktuálním členění věty, částečně intuicí, že jména a jejich zastupující zájmena se shodují v mnoha ohledech. Z dosažených výsledků úspěšnosti navržených atributů a ručně psaných pravidel je vidět, že se je třeba v budoucnu zvlášť věnovat studiu analýzy posesivní zájmenné anafory. U ručně psaných pravidel jsme použili zatím jen seznam kolokací textu, ve kterém se vyskytuje anafor. Dalším možným vylepšením může být kolokační slovník v rozsahu korpusu.

U demonstrativní zájmenné anafory jsme se nezaměřili na určení odkazujících ukazovacích zájmen, ani na určení zájmen, která odkazují k jmenným frázím nebo k větám (posloupnostem vět). Věnovali jsme se pouze ukazovacím zájmenům, která odkazují k jmenným frázím. Za nepříliš velkou úspěšností vyřešení této anafory může být fakt, že ke správnému určení antecedentů je potřeba znalosti širšího kontextu, protože ukazovací zájmena neodkazují jen k nejbližší jmenné frázi, která se shoduje v čísle a rodě, ale také k jmenné frázi v hlavní klauzi, po které ještě následuje vedlejší klauze. Nebo antecedent a ukazovací zájmeno se nemusí shodovat v čísle, protože ukazovací zájmeno v singuláru může odkazovat smyslově pouze k části antecedentu v množném čísle (např. Má tři klobouky. Ten bílý se jí líbí nejvíc.).

Pro určení antecedentů u nulové anafory kontroly a reciprocit jsme použili seznamy sloves a substantiv, vytvořené z větší části pozorování na trénovacích datech. Z tohoto důvodu nejsou seznamy úplné a jejich neúplnost může působit snížení úspěšnosti navržených pravidel na nových datech. S rozšířením korpusu o nová data bude tedy potřeba rozšířit i tyto seznamy.

Reflexivní a relativní a slovesné doplňky mají vysokou míru úspěšnosti díky své vlastnosti a struktuře PDT. Úspěšnost jmenných doplňků by se dala vylepšit použitím seznamu slovesných vazeb pro rozlišení řídicích jmen, která se s doplňky shodují v pádě nebo ne.

Úspěšnost určení antecedentů, závislého na informaci o čísle a rodě, byla ovlivněna také chybně označeným uzlům, které se vyskytly v datech. Absence pádu u nově vytvořených uzlů na tektogramatické rovině také měla vliv na výsledky určení jmenných doplňků.

V průběhu práce jsme narazili na problém absence lineárního pořadí na povrchové rovině u nově vytvořených uzlů na rovině tektogramatické. Předpokládáme, že vyřešení tohoto problému přispěje k přesnějším výsledkům.

* * * * *

Naše práce je jedna z mála prací věnovaných anaforám v češtině. Právě proto věříme, že přispěla k dalšímu vývoji výzkumu této problematiky.

Literatura

[Aone et Bennet, 1995] Aone, C. and Bennet, S.W. (1995). *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies*. In Proceedings of the International Workshop on Sharable Natural Language Resources (SNLR), 71-77, Nara, Japan.

[Hajičová et al., 1999] Hajičová, E., Panevová, J. a Sgall, P. (1999). *Manuál pro tektogramatické značkování*. ÚFAL Technical Report TR-1999-07, Univerzita Karlova, Praha.

[Hajičová et al., 2001] Hajičová, E., Havelka, J. a Sgall, P. (2001). *Discourse Semantics and the Saliency of Referents*. In Journal of Slavic Linguistics.

[Hajičová, 2003] Hajičová, E. (2003). *Aspects of Discourse Structure*. In Natural Language Processing between linguistic inquiry and system engineering, 47-54, Editura Universitatii Alexandru Ioan Cuza.

[Havránek et Jedlička, 1998] Havránek, B. a Jedlička, A. (1998). *Stručná mluvnice česká*. Fortuna, Praha.

[Jurafsky et Martin, 2000] Jurafsky, D. et Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall.

[Kučová et al., 2003] Kučová, L., Kolářová, V., Pajas, P., Žabokrtský, Z. a Čulo, O. (2003). *Anotování koreference v Pražském závislostním korpusu*. ÚFAL/CKL Technical report TR-2003-19, Univerzita Karlova, Praha.

[Kučová et Hajičová, 2004] Kučová, L. et Hajičová, E. (2004). *Coreferential Relations in the Prague Dependency Treebank*. Presented at 5th Discourse Anaphora and Anaphor Resolution Colloquium, San Miguel, Azores.

[Kučová et Žabokrtský, 2005] Kučová, L. a Žabokrtský, Z. (2005). *Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution*. In Matoušek, V.,

Mautner, P., Pavelka, T. (Eds.): Text, Speech and Dialogue, 8th International Conference, TSD 2005, Karlovy Vary.

[McCarthy et Lehnert, 1995] McCarthy, J.F. and Lehnert, W.G. (1995). *Using Decision Trees for Coreference Resolution*. In Proceedings of the 14th International Conference on artificial Intelligence (IJCAI-95), 1050-1055, Montreal, Canada.

[Mitkov, 2001] Mitkov, R. (2001). *Anaphora resolution*. Longman, London.

[Mooney, 2002] Mooney, R. (2002). Machine Learning. In Mitkov, R. (Ed.): *The Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford.

[PDT-průvodce] ÚFAL. *Pražský závislostní korpus 2.0* [online]. URL: <<http://ufal.mff.cuni.cz/pdt2.0/index-cz.html>>

[PDT-manuál] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razimová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K. a Žabokrtský, Z. *Anotace na tektogramatické rovině Pražského závislostního korpusu: Anotátorská příručka* [online]. URL: <<http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/pdf/t-man-cz.pdf>>

[Palek, 1988] Palek, B. (1988). *Referenční výstavba textu*. Univerzita Karlova, Praha.

[Panevová, 1991] Panevová, J. (1991). *Koreference gramatická nebo textová?* In Banys, W., Bednarczuk, L., Bogacki, K. (Eds.): *Etudes de linguistique romane et slave*, Krakow.

[Quinlan, 1993] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California.

[Sgall et al., 1980] Sgall, P., Hajičová, E., Buraňová, E. (1980). *Aktuální členění věty v češtině*. Academia, Praha.

[Soon et al., 2001] Soon, W.M., Ng, H.T. and Lim, C.Y. (2001). *A Machine Learning Approach to Coreference Resolution of Noun Phrases*. In *Computational Linguistics*, 27(4), 521-544.

Příloha A

Příklady anafory v dalších jazycích

*Poděkování patří mé milé kamarádce Evě Fialové,
která mi pomohla se zpestřením diplomové práce.*

- **(v angličtině):** Jane likes reading books. So does Marie. (Jana ráda čte knihy. Marie také Ø.)
- **(ve francouzštině):** Elle se lave chaque jour. (Ona se myje každý den.)
- **(v hov. holandštině):** Ik heb Piet z'n verstopen fiets gezien. (Viděl jsem „Petr jeho“ (= Petrovo) ukradené kolo.)
- **(v italštině):** Il mio figlio è molto pigro, Ø non fa i pulizi. (Můj syn je velmi líný. (on) Neuklízí.)
- **(v polštině):** Pani Kowalska wpadnie do sąsiadki, która teraz nie ma małżonka w domu. (Paní Kovalská se zastaví u sousedky, která teď nemá doma manžela.)
- **(v ruštině):** У мамы будет день рождения, а я пока не знаю, какой подарок ей купить. (Máma bude mít narozeniny, a ještě nevím, jaký dárek jí koupím.)
- **(ve španělštině):** A Juan lo veo. (Vidím „Juana ho“ (= Juana).)
- **(ve vietnamštině):** – Bao đi đâu rồi? (Kam šel Bao?) – Ø Đi học rồi. (Ø Šel do školy.) – ve vietnamštině se nulové zájmenné anafory používají jen při tykání, vynechání zájmen je považováno za nezdvořilé.

Příloha B

Seznam kontroly a reciprocity

Uvádíme zde seznamy sloves a substantiv kontroly, které byly používány v AČA.

V_ACT				
bát se	nabýt	podniknout	sloužit	vyvinout
bránit se	naklonit	pokoušet se	snažit se	vyžádat si
cítit	namáhat se	pokračovat	soustředit se	vznášet
cítit se	napadnout	pokusit	spěchat	vznést
časit	naskýtat se	pokusit se	stihnout	zabývat se
dařit se	naskytnout se	posbírat	stydět se	začínat
dát se	naučit se	potřebovat	svést	začít
dávat se	navázat	pouštět se	toužit	zahájit
dělat	navazovat	považovat	troufat si	zajít
dělat si	nechat se	pověst se	troufnout si	zamýšlet
dojít	nechávat se	provádět	učinit	zapomenout
dokázat	nést	provést	učit se	zapomínat
dopřát si	obávat se	přát si	udělat	zasloužit si
dostat se	obtěžovat se	předsevzít si	udělat si	zastávat
dovést	odcházet	přestat	ukázat se	zatožit
dovolit si	odejít	přestávat	ukazovat se	zaujímát
dovolovat si	odhodlat	přicházet	ukončit	zaujmut
dozvědět se	odhodlat se	přijet	umět	zavázat se
hodlat	odjet	přijímat	usilovat	zaznamenat
hrozit	odjíždět	přijít	uvažovat	zaznamenávat
chodit	odlétat	přijít si	uzavírat	zbýt
chtít	odmítat	přijíždět	uzavřít	zdráhat se
chtít se	odmítnout	přijmout	uznat	zdržet se
chystat se	odnaučit se	příslíbit	váhat	získávat
jet	odpovědět	příslušet	vědět	zkoušet
jevit	odvážit se	přistoupit	vejít	zkusit
jevit se	odvažovat se	přístupovat	vést	zkusit si
jezdit	opomenout	pustit se	vstoupit	ztrácet
jít	ostýchat se	rozhodnout	vstupovat	ztratit
konat	otálet	rozhodnout se	vyhýbat se	zůstat
koukat	plánovat	rozmyslit si	vykonávat	zůstávat
mínit	pocítit	rozpakovat se	vytknout si	zvládnout
mít	pocit'ovat	sbírat	využit	zvyknout si
moci	podářit se	slíbit	využívat	
nabízet se	podnikat	slibovat	vyvíjet	

V_ADDR				
bránit	naučit	pověřit	tlačit	vypomáhat
dojednat	navrhnout	pověřovat	učit	vypomoci
donutit	navrhovat	povolit	umožnit	vyzvat
dopomáhat	nutit	povolovat	umožňovat	vyzývat

dopomoci	obvinít	požádat	určit	zabránit
doporučit	obviňovat	požadovat	usnadnit	zabraňovat
doporučovat	oprávnit	předepsat	usvědčit	žádat
dovolit	opravňovat	přesvědčit	varovat	zakázat
dovolovat	otevírat	příkázat	velet	zakazovat
motivovat	otevřít	příkazovat	vést	zapovídat
nabádat	otvírat	přimět	vinit	zavázat
nabídnout	poděkovat	přinutit	vrátit	zavazovat
nabízet	podněcovat	přisuzovat	vybídnout	zmocnit
napomoci	pomáhat	radit	vybízet	znemožnit
nařídít	pomáhat	sloučit	vyčítat	znemožňovat
nařknout	pomoci	svádět		

V_PAT				
bavit	nutit	pozvat	škodit	vyplácet se
čekat	obžalovat	předurčovat	unavovat	vyplatit se
činit	osvědčovat se	připravit	určit	zadržet
jímat	osvobodit	snažit se	uvidět	zbýt
naučit	popadnout	spatřit	vadit	zbývat
nechat	postačit	stíhat	vidět	žalovat
nechávat				

V_BEN				
bývat	odmítat	označit	uznat	zdt se
jít	otevřít se	považovat	vyplatit se	znamenat
lze				

V_ORIG				
požadovat	vyžadovat			

V_ACT_ADDR				
dát	poskytnout	stavět	uložit	vyjádřit se
dávat	poskytovat	učinit	věnovat	vyjadřovat
klást	projevit	udělovat	vydat	vynést
podat	projevovat	udělit	vydávat	vysslovit
podávat	předat	ukládat	vyhlásit	vzdát
položít	předávat	ulevit	vyjádřit	

V_ACT_LOC				
budit	navozovat	vyvolávat	vzbudit	vzbuzovat
mít-respekt	vyvolat			

V_ACT_PAT	
najít	stačit

V_ACT_ORIG		
dostat	dostávat	získat

V_ACT_BEN_PAT	
být	

V_ACT_ADDR_BEN	
stát	

N_ACT				
cesta	nutnost	příjezd	spěch	volba
čas	ochota	příležitost	svoboda	vůle
cíl	odhodlání	připravenost	tendence	výtka
důvod	odvaha	problém	touha	zájem
hodláni	oprávnění	řešení	tradice	záměr
hrozba	plán	riziko	trend	záminka
chtíč	pokus	rozhodnutí	úkol	zásada
chuť	potřeba	rozhodování	umění	zásluha
jízda	povinnost	šance	úmysl	závazek
mínění	přání	schopnost	úsilí	zjednodušení
možnost	právo	síla	úvaha	zkouška
námaha	pravomoc	sklon	uznání	způsob
nárok	přestávka	slib	váhání	způsobilost
nechuť	příchod	snaha	vědomí	zvyk
neochota				

N_ADDR				
doporučení	pomoc	překážka	rozkaz	zákaz
motivace	povolení	příkaz	tlak	znemožnění
podezření	požadavek			

N_PAT			
osvědčení	příprava	škoda	únava

N_BEN				
odmítnutí	označení	prostředek	uznání	výhoda

n_ACT				
možnost naděje	pověření právo	přednost	příležitost	šance

n_ADDR				
doporučení nabídka	odpověď povolení	příslib rada	slib ujištění	zákaz zpráva

n_PAT				
pokuta				

n_ORIG				
návrh	souhlas			

Zde jsou seznamy reciprocity:

R_PAT_ADDR				
kombinace	propojit	sdužování	sjednocení	spojení
kombinovat	propojovat	sdužení	skloubit	spojit
kombinující	propojení	sjednocovat	sloučení	spojovat
komunikace	sdužit	sjednotit	sloučit	spojování
komunikovat	sdužovat			

R_ACT_EFF				
koordinovat	měření	měřit	srovnat	změřit

Příloha C

Ukázka C4.5

Ukážeme na příkladu reflexivní zájmenné anafory, jak funguje C4.5.

Vstupní soubor 1: refl.data, refl.test

- refl.data obsahuje vektory pro trénování, každý vektor je na zvláštním řádku
- refl.test obsahuje vektory pro testování

```
f,f,TWHEN,f,se-si,none
f,t,ACT,t,se-si,se-si
f,f,ACMP,f,se-si,none
f,f,PAT,t,se-si,none
f,t,ACT,t,se-si,se-si
f,f,ORIG,t,se-si,none
f,f,PAT,t,se-si,none
t,t,ACT,t,svuj,svuj
f,f,MEANS,f,svuj,none
```

Tab. C.5.2: Data na trénování

Vstupní soubor 2: refl.names

```
svuj, se-si, none. | classes

jedine zajmeno:      f, t.
i je subjekt:       f, t.
ifun:               ACT, AUTH, PAT, ADDR, EFF,
ORIG, ACMP, AIM, APP, BEN, CAUS, CNCS, COMPL,
COND, CONTRD, CPHR, CPR, CRIT, DENOM, DIFF,
DIR1, DIR2, DIR3, FPHR, HER, ID, INTT, LOC,
MANN, MAT, MEANS, PAR, PRED, REG, RESL, RESTR,
RSTR, SUBS, TFHL, TFRWH, THL, THO, TOWH, TPAR,
TSIN, TTILL, TWHEN, VOCAT, other.
i je aktant:        f, t.
jkat:               svuj, se-si.
```

Tab. C.5.3: Popis tříd a atributů

Výstupní soubor `ref1.dt` od C4.5

```

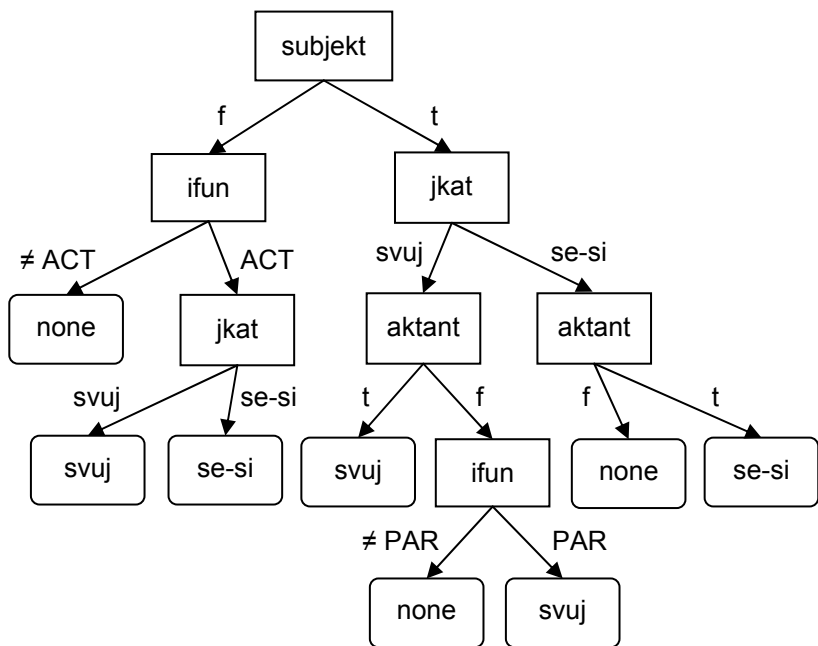
Simplified Decision Tree:

i je subjekt = f:
|   ifun = other: none (42.0/1.4)
|   ifun = ACT:
|       |   jkat = svuj: svuj (95.0/20.3)
|       |   jkat = se-si: se-si (58.0/16.9)
i je subjekt = t:
|   jkat = svuj:
|       |   i je aktant = t: svuj (2014.0/81.6)
|       |   i je aktant = f:
|       |       |   ifun = PAR: svuj (2.0/1.0)
|       |       |   ifun = other: none (0.0)
|       |   jkat = se-si:
|       |       |   i je aktant = f: none (16.0/2.5)
|       |       |   i je aktant = t: se-si (1154.0/60.8)

```

Tab. C.5.4: Ukázka výstupního souboru v C4.5

Rozhodovací strom



Obr. C.5.1: Zjednodušený diagram rozhodovacího stromu pro reflexivní zájmennou anaforu

Příloha D

Obsah CD-ROM

Součástí této diplomové práce je i přiložený CD-ROM, který obsahuje následující soubory a adresáře:

- `obsah.txt` – soubor s popisem obsahu CD-ROM
- `aca-diplomka.doc` – soubor s textem diplomové práce
- `/aca` – adresář obsahující implementace a vyhodnocení AČA