Anotování koreference v Pražském závislostním korpusu

Lucie Kučová, Veronika Kolářová,
Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo

Tato zpráva shrnuje výsledky práce na zachycování koreference v tektogramatických stromech Pražského závislostního korpusu dosažené v letech 2002–2003. Je zde především obsažen podrobný popis anotačního schématu, a to jak po stránce lingvistické, tak po stránce technické.

Výzkum se uskutečnil za finanční podpory grantu MŠMT LN00A063 a Německého akademického výměnného programu. Za četné konzultace děkujeme prof. Evě Hajičové, prof. Jarmile Panevové a prof. Petru Sgallovi.

Obsah

Kapitola 1

Úvod

1.1 Základní pojmy

Koreference je prvním z prostředků **koheze** – textové spojitosti vedle elipsy, substituce, lexikální koheze, konektorů a tematické posloupnosti ([Halliday and Hasanová, 1976]).

Reference je odkaz mluvčího k předmětům nebo situacím reálného světa. Rozlišuje se reference **exoforická** (poukazy k situaci, ke skutečnostem mimotextovým) a **endoforická** (výraz v textu poukazuje k jinému výrazu uvnitř téhož textu, má s ním shodnou referenci).[1] Jestliže se v textu vyskytují dva výrazy (nebo více výrazů) a poukazují k téže osobě, předmětu, skutečnosti – tj. jejich reference je identická – označuje se jejich vzájemný vztah, který propojuje výpovědi v textu, jako **koreference**.[2] Tento vztah určíme podle kontextu.[3]

Plnovýznamový výraz, k němuž se poukazuje, je běžně označován jako **antecedent**. Navazuje-li slovo na výraz v předchozí výpovědi nebo na předchozí výpověď jako celek, hovoříme o anaforickém navazování. Kataforické odkazování poukazuje na následující výpovědi nebo jejich části, zde se tedy analogicky jedná o **postcedent**. Odkazovat je možné v rámci věty, pak hovoříme o **intravětné** anafoře, nebo v rámci celého textu, pak jde o anaforu **intervětnou**.

Kromě těchto pojmů jsme zavedli pracovně také dvojici pojmů **koreferující** – **koreferovaný** [výraz]. Je obecnější a zanedbává prostorové umístění slov v textu – koreferovaným může být antecedent i postcedent.

Ve shodě s běžným přístupem české lingvistiky rozdělujeme koreferenci na **gramatickou** a **textovou**[4] a zpracováváme obě tyto oblasti anaforického odkazování.

1.2 Důvody pro zachycení koreference na tektogramatické rovině

V Pražském závislostním korpusu (*Prague Dependency Treebank*, PDT) je na rozsáhlý jazykový materiál aplikován Funkční generativní popis (FGD), vyvíjený pražskými lingvisty od konce šedesátých let. Důvody pro přítomnost gramatické koreference na tektogramatické rovině FGD jsou na příkladu německých vět ilustrovány v článku [Petkevič, 1995]:

[1] Používají se také pojmy **anafora** (vnitrotextové odkazování) a **deixe** (mimotextové odkazování).

[2] V zahraniční terminologii se setkáme s odlišením pojmů koreference a anafora. Jako **anafora** se označuje odkazování k nějaké (nejčastěji dříve zmíněné) jednotce v textu, kdežto **koreference** je odkaz mluvčího k totožným předmětům nebo situacím reálného světa. Tuto terminologii v našem pojetí zanedbáváme a řídíme se územ, který je běžný v české lingvistice.

[3] Pojem kontext je v tomto případě příliš mnohoznačný (i situace je kontext, stejně jako naše znalosti; v textu lze poukazovat i k širším kontextům politickým, kulturním aj.), proto se v případě endoforické reference hovoří někdy o **ko-textu** ([Hoffmannová, 1997]).

[4] K rozdělení koreference na gramatickou a textovou viz [Panevová, 1991].

Ich sah einen Politiker, welcher klug war.
**Ich sah die Magd, welcher klug war.*
Vztažné zájmeno ve vedlejší větě musí v rodu a čísle souhlasit se svým antecedentem (*Politiker* v první větě, *Magd* ve druhé větě), jinak nejde o správně utvořenou větu.

Jako zdroj dalších argumentů pro zachycování gramatické koreference můžeme použít konstrukce s doplňkem. Jednak některé typy doplňku rovněž vyžadují shodu s antecedentem,[5] např. *Honza přišel bos* v. *Marie přišla bosa*, jednak by při zanedbání koreference nebyl na tektogramatické rovině rozlišen rozdíl např. mezi větami *Poznal jsem ho ještě jako mladík* a *Poznal jsem ho ještě jako mladíka*.

Je tedy zřejmé, že gramatická koreference je nedílnou součástí popisu gramatických pravidel a intravětných vztahů, proto plně náleží na tektogramatickou rovinu FGD i PDT. Naproti tomu textová koreference sice není původní součástí návrhu FGD, neboť překračuje[6] úroveň jazykového významu a dostává se na rovinu smyslu, ale stává se tu vedle aktuálního členění další oblastí nadvětné lingvistiky, jejímž studiem se v rámci teorie diskursu PDT zabývá.

V průběhu anotace koreference v PDT vzniká nejen podrobný popis koreference v češtině,[7] ale také velký objem jazykového materiálu, použitelného k ověřování hypotéz a pro další lingvistický i aplikační výzkum vůbec.

1.3 Členění zbytku zprávy

V kapitole 2 zmíníme některé zahraniční přístupy k anotování koreference. Kapitola 3 popisuje anotační schéma zavedené pro koreferenci v PDT po technické stránce, kapitoly 4 a 5 pak obsahují pokyny pro anotaci jednotlivých typů gramatické a textové koreference. Základní kvantitativní vlastnosti dosud anotovaných dat a výsledky mezianotátorské shody jsou vyhodnoceny v kapitole 6.

[5]Někdy se v případě doplňku hovoří o dvojí závislosti.

[6]Ovšem i textová koreference je nutná k určení jazykového významu – např. homonymní věta *Předseda vlády řekl, že predloží návrh. . .* má jediné čtení až ve chvíli, kdy je určen antecedent nevyjádřeného subjektu vedlejší věty.

[7]Ačkoli v zahraničí je oblast anaforických vztahů již dlouho a obsáhle popisována, v prostředí české lingvistiky se se studiem koreferenčních vztahů (zejména u textové koreference) setkáváme zřídka.

Kapitola 2

Přehled zahraničních prací

2.1 Používané pojmy

Protože výzkum koreferenčních vztahů v zahraničí probíhá již dlouhou dobu, vytvořil se postupně poměrně rozsáhlý a ucelený pojmový systém. Je proto na místě se s ním seznámit, ačkoli se od něj česká terminologie poněkud odlišuje.[1]

Jak jsme se již dříve zmínili, v zahraničních pracích se nejčastěji setkáme s pojmem **anafora** (z řeckého $\alpha\nu\alpha\phi o\rho\alpha$, kde $\alpha\nu\alpha$ – zpět, nahoru; $\phi o\rho\alpha$ – nosit, vést). Anafora je zde striktně vázána na text, a to dokonce pouze na jednotlivý text; pokud jde o odkazování napříč texty, hovoří se zde o koreferenci (*cross-document coreference*). Z tohoto pojetí vyplývá, že je možné setkat se s textovými jednotkami, které jsou v koreferenčním vztahu, ale nejsou anaforické.

Rozlišuje se několik typů anafory. Základním termínem je **nominální anafora**: antecedentem referujícího výrazu (zájmena, substantiva nebo vlastního jména) je nezájmenná jmenná fráze. Z tohoto hlediska tedy můžeme rozlišit několik typů nominální anafory:

1. **zájmenná anafora** (zájmeno)
2. **určitá jmenná fráze** (substantivum s určitým členem)
3. **neurčitá jmenná fráze** (substantivum s neurčitým členem)
4. **one-anaphora** pro angličtinu (anaforický výraz je realizován výrazem „one“: *If you cannot attend a tutorial in the morning, you can go for an afternoon one*, srov. [Mitkov, 2001].[2]

Dalším z hledisek je rozdělení anafory na přímou a nepřímou.

- **přímá anafora** spojuje anaforu a antecedent na základě takových vztahů jako identita, synonymie či specializace. Antecedent je tedy s anaforou buď totožný, nebo je jí sémanticky blízký;
- v případě **nepřímé anafory** určujeme vztah mezi antecedentem a anaforou na pozadí nejen kontextu, ale zejména inference, která je založena na společném sdílení znalosti světa (*world knowledge*). Tento typ anafory se také označuje jako **asociativní** nebo **bridging anafora**. Pojítkem mezi antecedentem a anaforou bývá meronymie či holonymie.

[1] V celé této kapitole budeme pracovat s pojmy, které jsou jednotlivým přístupům vlastní, jakkoli neodpovídají terminologii FGD, která se používá v rámci PDT (to se týká zejména pojmů nominální/jmenná fráze).

[2] V češtině bychom nalezli paralelu tohoto typu anafory v případech adjektivního výrazu uvedeného bez substantiva, které je antecedentem.

Jestliže je antecedentem anafory více entit, které odkazují k totožnému referentu (jsou tedy v koreferenčním vztahu), označuje se jejich propojení jako **koreferenční řetězec** (*coreferential chain*).[3]

Většina systémů, které se studiem koreference zabývají, se soustředí pouze na ty případy, kdy je antecedentem jmenná fráze. Zpracování takového typu koreference, která odkazuje ke slovesným frázím, klauzím, větám, anebo dokonce k částem diskursu, je mnohem složitější, proto se s ním setkáváme jen zřídka.

Protože primárním přístupem k dané problematice je předpoklad, že každá nominální fráze (zejména na úrovni stejné nebo předchozí věty) je potenciálním antecedentem, zaměřují se jednotlivé systémy na definování množiny možných antecedentů, z níž se následně vybírají antecedenty reálné („správné“). K těmto úkolům slouží množství jednotlivých faktorů (*anaphora resolution factors*).

Ve většině přístupů převažují faktory „**vylučující**“ (*eliminating*, *constraints* – v různých terminologiích), které vyloučí určitou nominální frázi z množiny možných antecedentů. Do této kategorie řadíme zejména syntaktické metody. „Vylučující“ faktory jsou vždy obligatorní:

- morfologické aspekty (neshoda v rodě a čísle, příp. pádě);
- syntaktické aspekty (např. c-command);
- sémantická inkonzistence – antecedent musí být logický, tj. nemůže se neshodovat s povědomím o tom, jak reálný svět funguje.

Tyto aspekty bývají často kombinovány s faktory „**preferujícími**“ (*preferential*, *proposers*), které z množiny možných antecedentů některé nominální fráze upřednostňují. Protože nejde o faktory obligatorní, není jejich platnost aspektem určujícím, ale doprovodným:

- syntaktický paralelismus: za pravděpodobný antecedent se volí nominální fráze se stejnou syntaktickou rolí;
- sémantický paralelismus (u systémů, které identifikují sémantické role): vybrána je nominální fráze, která má stejnou sémantickou roli jako anaforický výraz;
- aspekty nadvětné syntaxe:
  - základ (*topic*) – ohnisko (*focus*): předpokládaným antecedentem je spíše *focus*;
  - vzdálenost mezi antecedentem a anaforou (*salience*): za pravděpodobný antecedent se volí nominální fráze, která je umístěna nejblíže koreferujícímu.

Současné přístupy k popisu anaforických jevů kombinují velké množství jednotlivých pravidel z různých oblastí lingvistiky od morfologie přes syntax, aktuální členění až po sémantiku, znalost světa a pragmatiku. Většina přístupů se soustředí na zájmennou nebo substantivní anaforu. Zpočátku převládala orientace na problematiku intravětné anafory, ale nyní jsou v centru zájmu jevy spojené s nadvětnou syntaxí, zejména se studiem bridging anafory. Různě se přistupuje k práci se znalostí světa – některé přístupy tuto oblast vědomě zanedbávají, jiné se na ni soustředí.

[3] V českém prostředí srov. také kohezní řetězec ([Kořenský et al., 1987]).

2.2 Anotační schémata

V oblasti koreference bylo vyvinuto několik anotačních schémat, která využívá, postupně rozpracovává a doplňuje řada lingvistických pracovišť. Pokládáme proto za nutné zde některá z nich zmínit a uvést jejich základní vlastnoti:

- Anaforický anotační systém UCREL[4]
  - využívá jej Lancaster Anaphoric Treebank;
  - široká oblast zpracování: od zájmenné anafory po elipsu a neanaforické použití zájmen;
  - anaforám jsou dodávány speciální symboly, které rozlišují směr reference (anaforická/kataforická) a označují různé sémantické charakteristiky anafory a antecedentu.
- Koreferenční anotační schéma MUC-7[5]
  - používá je mnoho pracovišť (University of Stendhal, Grenoble; Xerox Research Centre Europe; University of Wolvehampton aj.);
  - v textu identifikuje referující výrazy (substantiva, zájmena), které vstupují do koreferenčních vztahů, a označí typ vztahu mezi anaforou a antecedentem; jde tedy o zpracování koreferenčních řetězců;
  - zaměření na přímou anaforu.
- Schéma MATE[6]
  - anotování koreference v dialozích;
  - vychází ze schématu MUC, ale je obohaceno o mechanismy pro zaznačování dalších informací o anaforických vztazích;
  - vhodné zejména pro anaforické konstrukce typické pro románské jazyky (např. klitika);
  - umožňuje také vyznačování disambiguity.
- Schéma DRAMA
  - obdobné jako MUC, ale označuje i ty nominální fráze, které nevstupují do koreferenčních vztahů;
  - pokrývá širokou oblast referenčních výrazů: od zájmen (včetně jejich nulové formy) a substantiv po deixi[7] – deiktická adverbia, gesta a hlasové projevy;
  - soustředí se zejména na jevy spojené s bridging anaforou.
- Bruneseauxová & Romary[8]
  - anotační schéma pro dialogy mezi lidmi a přístroji;
  - pracuje se substantivy, „akcemi" (slovesa, klauze či celé věty) a s deiktickým odkazováním, které doprovází např. práce s počítačovou myší. Zahrnuje také referenci k vizuálnímu kontextu;
  - omezuje se pouze na ty referující výrazy, které jsou ve vzájemném vztahu.

---

[4] Srov. 2.3.2.
[5] Srov. 2.3.1.
[6] http://www.dfki.de/mate
[7] Deixi chápeme jako odkazování k mimotextovým skutečnostem.
[8] [Bruneseauxová and Romary, 1997]

2.3 Pracoviště zabývající se anaforickými jevy

V této podkapitole stručně přibližujeme několik projektů, které se zabývají popisem anaforických či koreferenčních vztahů v jiných jazycích než v češtině. Jak už bylo řečeno, v zahraničí je tato oblast lingvistiky v centru zájmu již dlouho a zabývá se jí – ať už teoreticky nebo na konkrétním jazykovém materiálu[9] – mnoho projektů a pracovišť. Popisujeme proto jen několik vybraných projektů,[10] abychom blíže ilustrovali přístupy, s nimiž je možné k dané problematice přistupovat. Zároveň uvádíme přehledový seznam nejznámějších projektů a pracovišť, které se anaforickými vztahy zabývají:

- **MATE**.[11] Rozsáhlý projekt, který se zabývá širokou škálou oblastí souvisejících s textovou lingvistikou. Koreferenční a anaforické vztahy zachycuje v dialozích.
- Projekty, které jsou součástí výzkumu koreferenčních vztahů Massima Poesia a Renaty Vieiry.[12] V obou fázích se zpracovávají texty z Wall Street Journal.
- **LORIA**.[13] Výzkum bridging anafory při francouzském korpusu PAROLE.
- **Secondary information structuring and comparative discourse analysis**.[14] Projekt Universität Bielefeld. Na materiálu vícejazyčného korpusu (japonština a jazyk kilivila) zpracovávají koreferenční vztahy v dialozích.
- **ANANAS**.[15] V tomto francouzském projektu jde o zachycování koreferenčních vztahů v různém typu textů: v textech odborných (administrativních a právních), publicistických i v beletrii. Součástí je také anotace koreference v dialozích.

Dále je anotace koreference součástí některých korpusů:

- **Penn Treebank**[16]
- **Potsdam Commentary Corpus**[17]
- **TIGER Corpus**[18]

Ve zbytku kapitoly probereme několik projektů podrobněji.

2.3.1 University of Wolverhampton

- na materiálu vlastního korpusu (obsahuje zejména technické manuály) zpracovává Research Group in Computational Linguistic referenčně identickou (přímou) nominální anaforu;
- hlavním úkolem je identifikace kompletních koreferenčních řetězců;
- přístup se tedy soustředí na určení referujících výrazů, které reprezentují první zmínky v textu o určité entitě, a následně na rozpoznání všech výrazů, které jsou s nimi v koreferenčním vztahu.[19]

[9] Nejčastěji je výzkum anaforických vztahů prováděn na částech již existujících korpusů.
[10] Ani tento seznam si neklade (z výše uvedených důvodů) za cíl podat vyčerpávající přehled.
[11] http://www.dfki.de/mate
[12] [Poesio and Vieira, 1998], [Poesio and Vieira, 2000]; přehledově viz http://www.cogsci.ed.ac.uk/ poesio/MATE/poesio1.html
[13] [Gardentová et al., 2002]
[14] http://www.text-technology.de
[15] Annotation Anaphorique pour l'Analyse Sémantique de Corpus; http://www.inalf.fr/ananas
[16] http://www.cis.upenn.edu
[17] http://www.ling.uni-potsdam/de/cl/cl/les/frosch_fce.html
[18] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus; více viz http://www.coli.uni-sb.de/ hansen/kunz-hansen.ps
[19] Daná procedura se zakládá na předpokladu, že koreferenční vztah je „tranzitivní" – každý prvek koreferenčního řetězce je označen jako identický s původním (nejdříve zmíněným) výrazem.

Anotování je založeno na schématu MUC-7 Coreference Task Definition,[20] v němž je kladen větší důraz na vysoký stupeň mezianotátorské shody než na to, aby byly pokryty všechny typy koreferenčních vztahů. Zaznačuje pouze vztah identity mezi dvěma substantivy (stranou zůstávají odkazy ke klauzi, diskursu či bridging anafora).[21]

Anotační schéma rozlišuje dva základní typy označení:

- první zmínka o referující entitě v textu (prvním bodě koreferenčního řetězce) má symbol <COREF ID="###">; ID je identifikačním číslem entity (Každý člen koreferenčního řetězce má svoje vlastní ID);
- ostatní členy koreferenčního řetězce jsou označeny <COREF ID="#" TYPE="IDENT" REF="#">.

Atribut TYPE má vždy hodnotu IDENT (vztah mezi entitami je identita). Atribut REF vypovídá o tom, se kterou entitou je dané slovo v koreferenčním vztahu - hodnotou REF je tedy ID první položky koreferenčního řetězce.

```
 <COREF ID="2">The illustration on <COREF ID="3">the facing
page</COREF></COREF> shows <COREF ID="4">all the equipment you
will need to set up <COREF ID="5" TYPE="IDENT" REF="1">your
computer</COREF> and begin using <COREF ID="6" TYPE="IDENT"
REF="1">it</COREF></COREF>. Place <COREF ID="7" TYPE="IDENT"
REF="1">your equipment</COREF> on <COREF ID="8">a sturdy, flat
surface</COREF> near <COREF ID="9">a grounded wall outlet</COREF>
Before following <COREFID="10">the setup instructions in <COREF
ID="11">this chapter</COREF></COREF>, you may want to read<COREF
ID="12">"Arranging your Office " in <COREF ID="13">Appendix
A</COREF>, in <COREFID="14">the section on <COREF
ID="15">health-related information</COREF></COREF></COREF>,
for<COREF ID="16">tips on adjusting <COREF ID="17">your work
furniture</COREF></COREF> so that you're comfortable when using
<COREF ID="18" TYPE="IDENT" REF="1">the computer</COREF>.
```

Obrázek 2.1: Ukázka anotace podle University of Wolverhampton.

2.3.2 Lancaster Anaphoric Treebank

Lancaster University, University Centre for Computer Corpus Research on Language (UCREL)[22]

- anotováno cca 100 000 slov z Associated Press Treebank Corpus, částečně také ze Spoken English Corpus (rozhlas);
- schéma[23] je založeno na tom, že v textu identifikuje referující výrazy, mezi nimiž vyznačí jednotlivé vztahy (jejich směr i typ);
- zaměření: zejména na jevy spojené s bridging anaforou.

[20][Hirschman, 1997]

[21]Pracoviště používá vlastní anotační prostředek ClinkA ([Orăsan, 2000]), který rozvíjí MUC-7 Coreference Annotation Task a je využíván jako jeho doplněk.

[22]Součást korpusu Associated Press Treebank (AP); http://www.comp.lancs.ac.uk./computing./research/ucrel

[23][Fligelstone, 1990]

Označují se:

- substantiva
- zájmena (včetně posesiv)
- části vět, kde je nutné doplnit elidované sloveso
- klauze
- „pro-verbs“ (*do*, *do so*)

Směr:

- $<$ anafora
- $>$ katafora
- $<>$ směr je nejasný nebo je na výběr z více možností
- $><$ exofora

Typy vztahů:

- **REF=** koreference: zájmeno
- **(1 1), (1 1)** koreference: substantivum
- **SUBS=** substituce (*more changes than those reccomended by the company was looking for a taxpayer group. Finding none. . .*)
- **ELLIP=** elipsa (*the UK isn't a republic. If it were, . . .*)
- **IMP=** určitá nepřímá anafora – implikovaný antecedent (*The 'love triangle' case... the defendant*)
- **OF=** substantivum s předpokládaným doplněním *of* (*Edinbourgh High School. . . the headmaster*)
- **MISC=** různé typy bridging anafory (množina/podmnožina, část/celek)
- {{. . . . . .} apozice u substantiv (*the prime minister, Mr. Blair*)
- **MTR=** metatextová koreference

Každý nezájmenný referující výraz je v závorce a je označen identifikačním číslem. Koreference je znázorněna tím, že referující výraz dostane stejné identifikační číslo jako jeho antecedent.

```
 (1 The deputy leader of (2 the Labour-controlled Liverpool
Council 2) 1), {{1 Mr Derek Hatton 1}, has said any of (3
<REF=1 his town hall staff 3) who (4 cooperate with (5 the
district auditor 5) 4) faced the threat of suspension. Yesterday,
<MISC=3(6 two senior officials who'd <SUBS=4:3done just that
6) were sent home, pending an investigation. (5 The auditors 5)
were called in immediately after (5 the council 5) agreed to a
nine per cent rates increase to fund spending beyond Government
limits. (1 Hr. Hatton 1) said last night <REF=1 he didn't know?(7
whether <MISC=3,6(8 officers 8) were obliged by law to cooperate
with (5 the district auditor 5) 7)?. He added: ' <REF=1[S]I'm not
bothered about <REF=7 that'.
```

Obrázek 2.2: Ukázka anotovaného textu z Lancaster Anaphoric Treebank.

2.3.3 Xerox Research Centre Europe ve spolupráci s University of Stendhal

- anotace korpusu o velikosti zhruba milion slov;[24]
- zaměření spíše na nadvětné vztahy než na čistě syntaktické intravětné jevy, proto se nepracuje např. s reflexivními zájmeny (která vždy koreferují se subjektem) či se vztažnými zájmeny;
- orientace na nejbližší antecedenty anaforických výrazů, ne na celé koreferenční řetězce;
- na rozdíl od projektů MATE a MUC, které vnímají koreferenční vztahy jako symetrické, je zde akcentováno směřování od anafory k antecedentu.

Zpracovává se:[25]

- osobní zájmena ve 3. osobě
- přivlastňovací zájmena
- ukazovací zájmena kromě tzv. *neuter pronouns* (*ce*, *ça*, *cela*, *ceci* – to, toto, tamto)
- neurčitá zájmena, číslovky
- adjektiva
- idiom *le* + *faire* (spojení zájmena a slovesa = udělat [to])
- anaforická adverbia jako *dedans*, *dessus* (v, na)
- substantivní elipsy
- anaforické odkazy označující ohraničenost, uzavřenost, např. *ce dernier*, *le premier* (tento poslední/první)

Anotace zachycuje nejen informaci o prvcích zahrnutých v anaforických vztazích (tedy to, který anaforický výraz je vztažen ke kterému antecedentu), ale také charakterizuje vztah mezi anaforou a antecedentem.

Mezi anaforou a antecedentem se rozlišují tyto vztahy:

- **koreference** – anaforický výraz odkazuje k referenčně totožnému antecedentu;
- **set-membership** – vztah je založen na souvislosti část-celek, množina-podmnožina; antecedent je množina, anaforický výraz může být její podmnožina nebo jednotlivý prvek;
- **deskripce** – ani anaforický výraz, ani jeho antecedent referenta neoznačují, pouze jej popisují. Do této kategorie spadají především klitika;
- **phrase** – antecedentem je klauze nebo věta;
- **neurčitý vztah** – pro ty případy, které nepokryly předchozí čtyři typy (např. u negativního kvantitativního určení: *Among these students, none has done his work.*).

Popis anotačních značek:

- <**p**> – vymezuje odstavec;
- <**s**> – signalizuje hranice věty;
- <**exp**> – výraz, který je buď anaforický, nebo je antecedentem anaforického výrazu. Každý má vlastní atribut id, který je jedinečný a identifikuje jej v dokumentu;
- **next, prev** – atributy, které slouží při označení jednotlivých částí „roztrženého“ anaforického výrazu (např. při vložení vysvětlující klauze). Celý anaforický výraz je označen dvěma symboly <exp>; první <exp> je doplněn atributem next=”X” (X je id druhé části anaforického výrazu), druhý <exp> má atribut prev=”Y” (Y je id první části anafory);

[24] [Tutinová et al., 2000]

[25] Antecedent není omezen pouze na substantivum, může jím být i klauze či věta.

- <**ptr**> – prázdný element, který označuje místo v textu, kde se objevil referující výraz (doplňuje se automaticky);
- **src** – označuje antecedent; může mít několik hodnot (pokud pro anaforický výraz existuje více antecedentů);
- <**ptr-i**> – pro případy, kdy nelze antecedent určit jednoznačně.

Typ vztahu mezi anaforou a jejím antecedentem (případně nulová existence takového vztahu) se označuje atributem „type“ u elementu <ptr>. Hodnota tohoto atributu tedy může být:

- **coref** – koreference
- **mde** – set-membership
- **desc** – deskripce
- **phrase** – větný antecedent
- **indef** – neurčitý vztah

Existují i další atributy, např. atribut „coord“ pro koordinačně připojené antecedenty.

```
 <p n="78" id="PO78">
<s> <exp id="e131">L'expression oeuvre scientifique</exp>,
objet de notre étude ne se laisse pas facilement appréhender
par le droit. </s><s> On peut <exp id="e132"><ptr type="coref"
src="e131"/>lui</exp> donner un sens très général et considérer
que l'expression vise toute production intellectuelle de
caractère scientifique (§ 1. ). </s><s> Il est possible de <exp
id="e133"><ptr type="coref" src="e131"/>lui</exp> donner un
contenu plus restreint si l'on met l'accent sur le terme oeuvre
(§ 2. ).</s></p> <p n="79" id="PO79"> <s> <exp id="e134">Le mot
oeuvre</exp> a, en droit, plusieurs significations. </s><s>
Selon le vocabulaire juridique de l'association Henri Capitant
dirigé par le Doyen Cornu, <exp id="e135"><ptr type="coref"
src="e134"/>il</exp> revêt notamment les sens suivants : </s>
</p> <p n="80" id="PO80"> <s> ouvrage résultant d'une
construction (immobilière) ; </s>
</p> <p n="81" id="PO81"> <s> activités déployées en vue d'un but
déterminé (activités de l'entreprise ou activités universitaires
et sociales). </s> </p> <p n="82" id="PO82"> <s> D'une manière
générale, <exp id="e136"><ptr type="coref" src="e134"/>il</exp>
s'analyse comme le résultat d'un travail ou d'une activité
manuelle ou intellectuelle. </s><s> À l'évidence, c'est cette
dernière acception qui semble la plus adaptée pour notre étude.
</s>
```

Obrázek 2.3: Ukázka anotace podle Xerox Research Centre Europe.

Kapitola 3

Technické provedení v PDT

3.1 Datová reprezentace

3.1.1 Předcházející návrhy

V následujících odstavcích odkážeme na dva předcházející návrhy, jak v tektogramatických stromech reprezentovat koreferenci.

Formální zachycení gramatické koreference navržené v [Plátek et al., 1984] (v následujícím popisu čerpáme z [Petkevič, 1995]) vychází z předpokladu, že ve stromu je nutné označit tzv. relativní cestu, která vede od antecedentu ke koreferujícímu zájmenu. K tomuto účelu je u uzlů zaveden atribut `relpath` s následujícími hodnotami:

- `RELSTART` – označuje počáteční uzel relativní cesty;
- `RELCONT` – relativní cesta pokračuje přes takto označený uzel, nebo v něm končí;
- `0` – přes uzel nevede žádná cesta.

V rámci anotačního schématu PDT byl původně navržen další typ reprezentace koreference ([Hajičová et al., 2000], str. 6 a str. 44–45, [Hajičová et al., 2000]), podle kterého měl mít každý uzel čtyři atributy:

- `ANTEC` – funktor antecedentu u gramatické koreference, nebo `NIL`;
- `COREF` – lemma antecedentu koreference, nebo `NIL`;
- `CORNUM` – číslo antecedentu (pořadové číslo uzlu ve stromu);
- `CORSNT` – věta, ve které je antecedent; atribut má možné hodnoty `NIL` (antecedent je v rozebírané větě) a `PREV` (antecedent je v předcházející větě).

3.1.2 Současné řešení

Současné řešení využívá skutečnosti, že každý uzel každého tektogramatického stromu má identifikátor, který je jedinečný v celém PDT a který je uložený v atributu TID. Jestliže můžeme chápat koreferenci jako odkaz z jednoho uzlu na jiný uzel, pak stačí do vybraného atributu počátečního uzlu uložit identifikátor cílového uzlu. Ke každému takovému odkazu je ještě třeba určit typ koreference, kterou zastupuje (gramatická/textová).

K zachycení koreference byly u každého uzlu tektogramatických stromů[1] zavedeny čtyři nové atributy:

- `coref` – identifikátor cílového uzlu (nebo posloupnost takových identifikátorů oddělených znakem |, vede-li z uzlu více odkazů);[2]
- `cortype` – podle typu koreference hodnota `gram`, `text` a `comp`[3] (nebo posloupnost typů oddělených znakem |, vede-li z uzlu více odkazů; typy odkazů jsou seřazeny ve stejném pořadí jako identifikátory v atributu `coref`);
- `corlemma` – buď lemma entity, na kterou tento uzel odkazuje, ale která není reprezentována žádným uzlem v okolních stromech, nebo jiná speciální hodnota:
  - `segm` - viz kapitola 5.1.2
  - `exoph` - viz kapitola 5.1.3
- `corinfo` – pracovní komentář, který upozorňuje na
  - na případy, kdy je struktura v rozporu se sémantikou (např. nepravé vedlejší věty vztažné, srov. kapitola 4.2.4);
  - na konstrukce, kde anotátor nedokáže antecedent určit.

Pokud uzel není koreferenční, obsahují všechny tyto atributy prázdný řetězec.

Jestliže vede koreferenční vztah k uzlu, který v tektogramatickém struktuře není listem, implicitně předpokládáme, že antecedentem je nikoli tento jediný uzel, ale celý podstrom, který kromě daného uzlu obsahuje i všechny jeho potomky. Například ve větě *Můj o dva roky mladší bratr, kterého ještě neznáš, přijde zítra* koreferuje vztažné zájmeno vedlejší věty s výrazem *Můj o dva roky mladší bratr*, nikoli jen s výrazem *bratr*.[4]

Speciálním případem koreference s nelistovým uzlem je odkazování k celé větě (např. kapitola 5.1.1). V takové situaci označujeme za antecedent hlavní predikát věty, nikoli technický kořen tektogramatické struktury s funktorem SENT.

3.2 Anotační rozhraní – editor stromů Tred

Anotování koreference v souborech tektogramatických stromů probíhá, stejně jako většina ostatních anotací v PDT, v editoru stromů Tred.[5] Aby byla ruční anotace koreference maximálně efektivní a co nejméně náchylná ke vzniku chyb způsobených nepozorností, bylo vhodné vytvořit v Tredu nový režim (kontext) `Coref` se specifickými vlastnostmi, viz obr. 3.1. Především jde o

- nový způsob vizualizace koreferenčních vztahů – koreference mezi dvěma tektogramatickými uzly je v Tredu v režimu `Coref` zobrazována jako oblá přerušovaná šipka, která vede od počátečního (odkazujícího) uzlu k cílovému uzlu (antecedentu), obvykle tedy zprava doleva, viz obrázek 3.1; gramatická a textová koreference jsou odlišeny barvou;[6]

[1] Soubory tektogramatických stromů jsou během anotace uloženy ve fs-formátu, vyvinutém pro potřeby PDT. Každý soubor obsahuje zhruba padesát vět-stromů. Na věty z PDT v této zprávě odkazujeme podle vzoru `ml19#21` (název souboru#číslo věty v něm).

[2] Např. antecedentem osobního zájmena ve větě *Marie vzala Vlastu do divadla, kde na ně čekal Marek* jsou dva uzly, ke kterým je nutno odkázat jednotlivě.

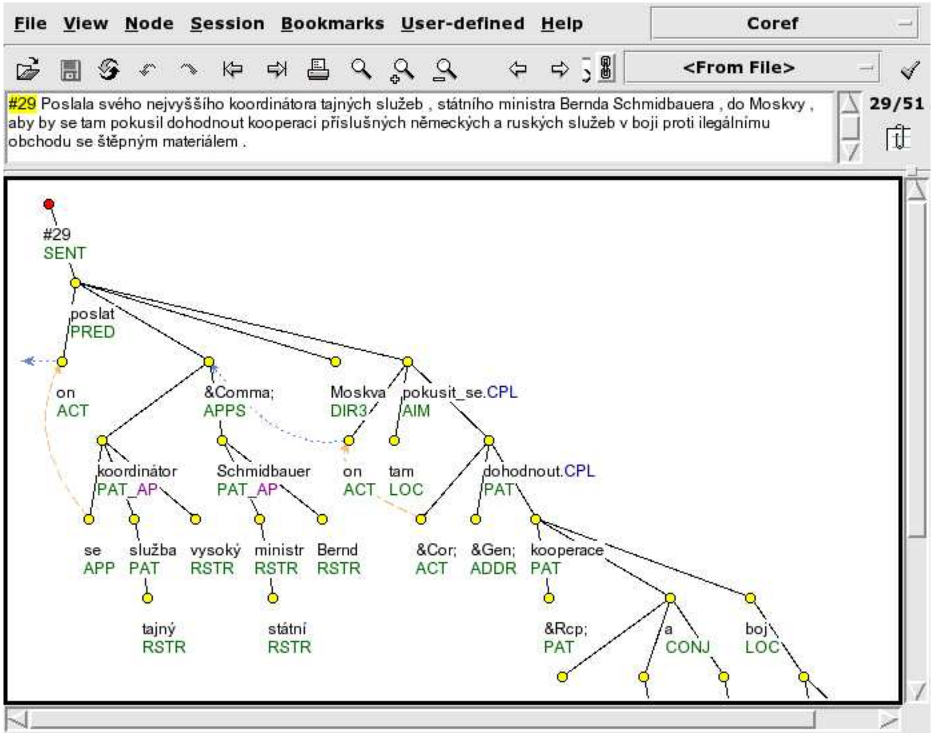
[3] V případě hodnoty `comp` nejde ve skutečnosti o zachycení koreference, ale o druhou závislost u doplňku (viz kapitola 4.3). Tato hodnota zatím nebyla použita (místo ní je v datech hodnota `gram`), ale bude zanášena zpětně.

[4] Nelze vyloučit existenci případů, kdy je možné za antecedent považovat právě jen kořen příslušného podstromu, ale ne už jeho potomky (nebo kořen a část jeho potomků, ale ne všechny). V navrženém anotačním schématu tuto možnost zatím nereflektujeme.

[5] http://ckl.mff.cuni.cz/pajas/tred

[6] V této zprávě je šipka odpovídající gramatické koreferenci znázorněna přerušovanou čarou a šipka odpovídající textové koreferenci tečkovaně.

- rychlé a pohodlné ovládání – uvnitř jednoho stromu lze koreference snadno „kreslit“ pomocí myši, pro zaznamenávání koreference přes hranice vět jsou zavedeny klávesové zkratky.



Obrázek 3.1: Editor stromů Tred v režimu `Coref`.

3.3 Pokusy s částečnou automatizací

3.3.1 Zvýraznění „kandidátů“

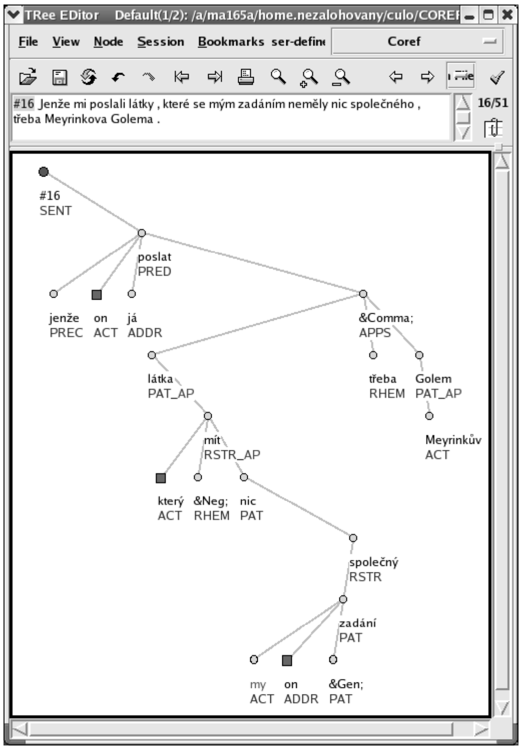
Nejjednodušší způsob, jak usnadnit anotátorům práci, je automatické rozeznávání koreferujících uzlů (zatím bez určení antecedentu). V Perlu byla implementována funkce, která tyto uzly rozezná s vysokou úspěšností. Uzel, který tato funkce označí jako koreferující, je v Tredu v režimu `Coref` zvýrazněn jinou barvou a větší velikostí uzlu (viz obr. 3.2), a to až do okamžiku, kdy anotátor sám určí antecedent. Anotátor tak nalezne koreferující uzly daleko rychleji (i když o antecedentech už se musí rozhodnout sám).

Ze 10 739 koreferujících uzlů, které byly obsaženy v testovacím vzorku 269 souborů (více o anotovaných datech viz kapitola 6), jich bylo zmíněnou funkcí správně rozeznáno 10 555 (98,3 %). Funkce nesprávně označila pouze 184 ve skutečnosti nekoreferujících uzlů („nadgenerování“ 1,7 %).

3.3.2 Určování antecedentů při gramatické koreferenci

Dalším krokem po hledání koreferujících uzlů je přirozeně hledání jejich antecedentů. Při pokusech s automatickou anotací jsme se zaměřili pouze na gramatickou koreferenci. Bylo navrženo a otestováno několik ručně psaných pravidel:[7]

[7]Zde uvádíme pouze ta z testovaných pravidel, která dosáhla přijatelné úspěšnosti.



Obrázek 3.2: Uzly „kandidující“ na koreferenci jsou odlišeny tvarem a barvou.

- `RelativeClauseRule` – nad uzlem s lemmaty *který*, *jenž*, *jak* najdi nejbližší uzel s funktorem RSTR (kořen vztažné věty); substantivum nebo zájmeno nad ním vyber jako antecedent;
- `ReflexiveRule` – nad uzlem s lemmaty *se*, *svůj* najdi nejbližší uzel, který má mezi potomky uzel s funktorem ACT a nemá lemma *se*; tento potomek vyber jako antecedent;
- `ControlRuleACT`, `ControlRulePAT`, `ControlRuleADDR` – je-li nad slovesem v infinitivu ve stromě další sloveso, pak jde většinou o tzv. sloveso kontroly. V takovém případě koreferuje nevyjádřený subjekt infinitivu s jedním z aktantů řídícího slovesa.[8] O který aktant jde, je závislé na řídícím slovesu (*nutit* - ADDR, *snažit se* - ACT . . .). Seznam sloves kontroly byl získán ze slovníku VALLEX 1.0.[9]

Při vyhodnocování úspěšnosti jednotlivých pravidel (viz tabulka 3.1) byly použity následující pojmy:

- *Cover* - počet uzlů, kde dané pravidlo určilo antecedent, děleno počet uzlů, u kterých má být antecedent určen;
- *Recall* - počet uzlů, kde dané pravidlo určilo antecedent správně, děleno počtem uzlů, u kterých má být antecedent určen;
- *Precision* - počet uzlů se správně určeným antecedentem děleno počtem uzlů s určeným antecedentem.

Automatickou proceduru mohou anotátoři spouštět přímo v Tredu jednotlivě nad každým stromem.

[8] Výjimečně může koreferovat i s některým volným doplněním, např. BEN nebo LOC (více viz kapitola 4.4).
[9] http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/

Rule	*Cover*	*Recall*	*Precision*
`RelativeClause`	1751 (33.88 %)	1671 (32.33 %)	95.43 %
`ControlADDR`	44 (0.85 %)	39 (0.75 %)	88.64 %
`Reflexive`	1122 (21.71 %)	979 (18.94 %)	87.25 %
`ControlACT`	675 (13.06 %)	472 (9.13 %)	69.93 %
`ControlPAT`	15 (0.29 %)	5 (0.10 %)	33.33 %
Celkem	3607 (69.8 %)	3166 (61.2 %)	87.8 %

Tabulka 3.1: Vyhodnocení úspěšnosti pravidel pro automatickou anotaci gramatické koreference. V testovacím vzorku bylo celkem 5168 uzlů s gramatickou koreferencí (100 % pro *cover* a *recall*).

Kapitola 4

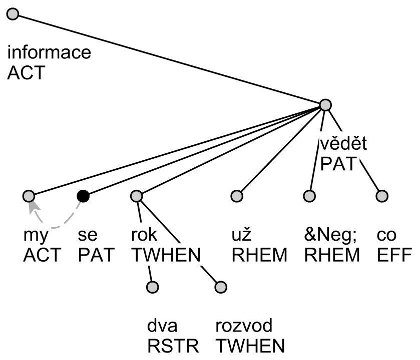
Gramatická koreference

Gramatická koreference je určována gramatickými pravidly, na jejichž základě je zpravidla možné určit antecedent.[1] Gramatická koreference je realizována následujícími jazykovými prostředky a syntaktickými konstrukcemi:

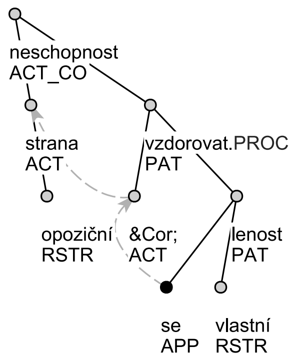
- zvratná zájmena (kapitola 4.1)
- vztažné prostředky (kapitola 4.2)
- doplněk (kapitola 4.3)
- kontrola (kapitola 4.4)

4.1 Zvratná zájmena

Reflexiva (tj. zvratná zájmena osobní i přivlastňovací, např. obr. 4.1 a 4.2) koreferují v pozicích, kdy jsou pokládána za větný člen. Mají společné trlemma *se*. Poukazují primárně k subjektu věty, a to k nejbližšímu Aktorovi (tedy primárně k Aktorovi stejné klauze; pokud není Aktor ve stejné klauzi, koreferuje *se* s Aktorem řídící klauze/věty).



Obrázek 4.1: Zvratná zájmena: *Informace o tom, co o* ***sobě****, dva roky po rozvodu, už nevíme.* (lm40#2)



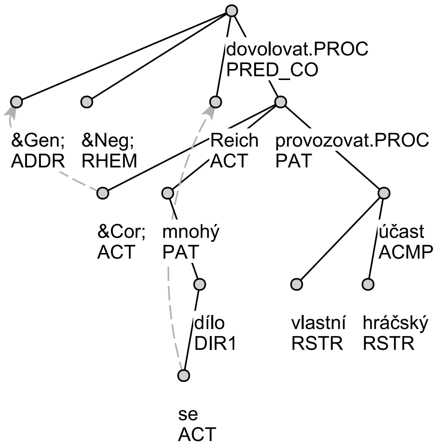
Obrázek 4.2: Zvratná zájmena: *Neschopnost opozičních stran vzdorovat* ***své*** *vlastní lenosti.* (lm39#16)

[1] Vždy však nemusí platit, že je antecedent možné stanovit na základě gramatických pravidel jednoznačně. Například ve větě *Přinesl jsem ti knihu o Boženě Němcové, kterou mám velice rád* jím může být stejně tak *Božena Němcová* jako *kniha*. Tato možná nejednoznačnost je však ošetřena strukturou věty na tektogramatické úrovni: ze stromové struktury je zřejmé, které slovo je řídícím uzlem vedlejší věty.

Antecedentem **osobního zájmena *se*** je v aktivních větách takřka vždy Aktor klauze (věty), v pasivních větách je antecedentem příslušný aktant valenčního rámce slovesa.

Mnohem složitější je situace u **posesiva *svůj***. I zde sice můžeme říci, že *svůj* primárně (a nejčastěji) koreferuje s nejbližším Aktorem, ale můžeme se setkat také s jinými případy:

- u sloves kontroly,[2] kde je kontrolujícím členem jiný aktant než Aktor[3]. Zde nekoreferuje *svůj* s nejbližším Aktorem (tedy Aktorem kontrolovaného infinitivu/nominalizace s lemmatem Cor.ACT), ale s Aktorem slovesa kontroly, viz obr. 4.3.
- ve větách se slovesem ve 3. osobě. Zde se *svůj* používá pro koreferenci ke kterémukoli aktantu (např. obr. 4.4). Platí to zejména v případech nestandardního použití zájmena *svůj*. O netypickém použití zájmena *svůj* vypovídá často už to, že koreferovaným je aktant s jiným funktorem než ACT.



Obrázek 4.3: Zvratná zájmena: *Mnohá ze **svých** děl Reich nedovoluje provozovat bez vlastní hráčské účasti.* (lr19#38)

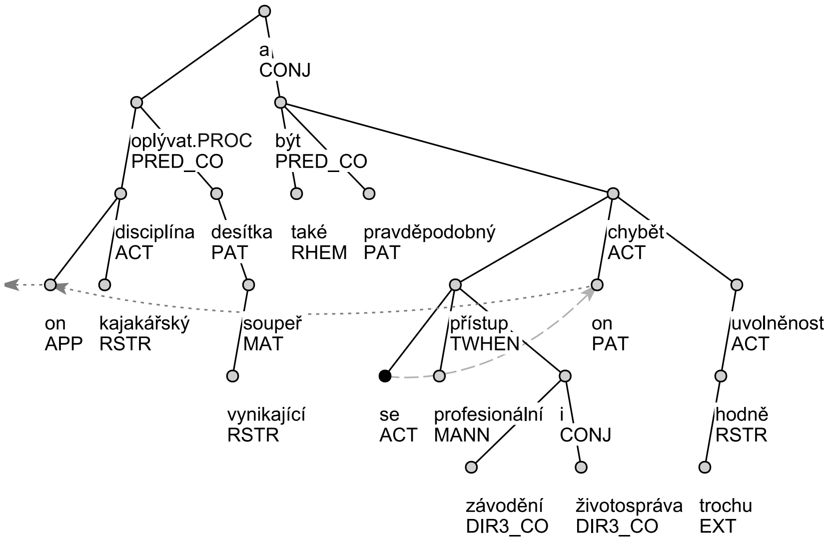
Také záměna zájmena *svůj* za zájmeno *jeho* se objevuje stále častěji (srov. např. [Čmejrková, 1998], příklad viz obr. 4.5). Respektujeme záměr mluvčího, zájmeno ponecháváme v dané podobě a zaznačíme příslušný koreferenční vztah. Informaci o nestandardním použití zájmena zachováváme v atributu `corinfo`.

V současnosti se stále více setkáváme také s posesivním zájmenem v příslušné osobě u 1. a 2. osoby, zejména v jazyce reklamy a v žurnalistice (např. *Užijte si **vaši** dovolenou!*). I v těchto případech respektujeme záměr mluvčího, zájmeno ponecháváme v dané podobě a podle použité formy volíme druh koreference. Také zde uvedeme informaci o nestandardním použití zájmena v atributu `corinfo`.

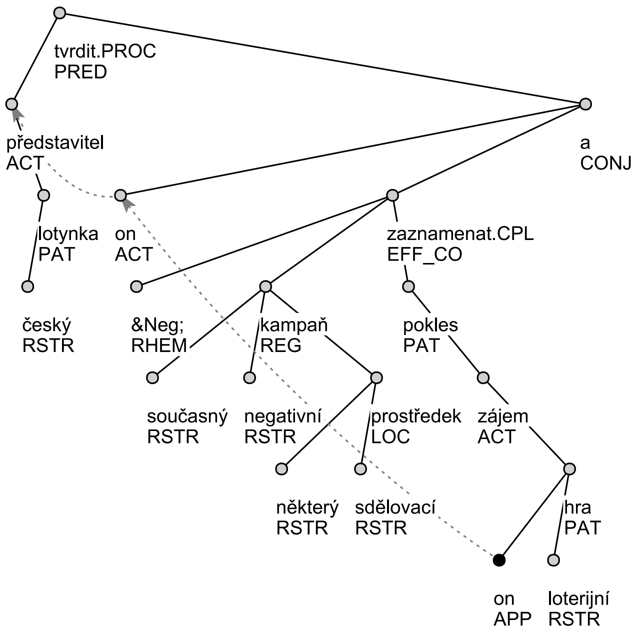
U zvratných zájmen nezaznačujeme koreferenci ve dvou případech:

1. O koreferenční vztah se nejedná v případech, kdy *se* tvoří součást **ustáleného spojení** nebo je samo **frazémem**. Takové použití trlemmatu *se* většinou signalizuje funktor DPHR (*dependent part of phraseme*), ale anotátoři koreference berou v úvahu také systémově neoznačená idiomatická vyjádření.
   Prozatímní seznam idiomatických vyjádření:
   ▷ *Cena akcií vzrostla i u IPB, což je **svým způsobem** podivuhodné.*
   ▷ *Knihy dejte na **svá** místa.*

[2]Srov. kapitola 4.4.
[3]K problematice vnitřní predikace viz [Panevová, 1986].



Obrázek 4.4: Zvratná zájmena: *Jejich kajakářské disciplíny oplývají desítkami vynikajících soupeřů a je také pravděpodobné, že při **svém** profesionálním přístupu k závodění i k životosprávě jim chybí trochu víc uvolněnosti.* (ml11#3)



Obrázek 4.5: Zvratná zájmena: *Představitelé České lotynky tvrdí, že v souvislosti se současnou negativní kampaní v některých sdělovacích prostředcích nezaznamenali pokles zájmu o **jejich** loterijní hry. . .* (ls39#11)

- ▷ *Podivínský premiérův nápad kritiku stimuluje **sám o sobě**.*
- ▷ *Jít si **po svých**.*
- ▷ *Začali pošesté, tentokrát **na svém**.*
- ▷ *Ve **své době** to byl provokativní krok.*
- ▷ ***Stál** si pevně **na svém**.*
- ▷ *Dnes **jsme** již sedm let **svoji** a domníváme se, že jsme si čím dál tím bližší.*
- ▷ ***Jsme sví**.* (tj. zvláštní, jiní)
- ▷ *Pošestnácté **za sebou** se na dvorcích ve Flushing Meadow představí Ivan Lendl.*
- ▷ *Délka nájmu je stanovena na 20 let s možností dalšího pronájmu na 10 let, a to třikrát **po sobě**.*
- ▷ *Stupeň jakosti Q označuje na rozdíl od značky Czech Made světově proslulou jakost srovnatelnou se špičkovými výrobky **svého** druhu.*
- ▷ *Americký ministr získal už předem přízeň Pekingu, který jej za deklarovaný důraz na hospodářské aspekty zahraniční politiky **sám od sebe** protokolárně povýšil na prezidentského vyslance.*

2. O koreferenci neuvažujeme ani u tzv. **etického dativu** (ETHD):
   - ▷ *Potentáti v bance koupí za deset, prodají **si** za patnáct.*

4.2 Vztažné prostředky

4.2.1 Vedlejší věty vztažné

Vztažná zájmena i zájmenná příslovce uvádějící vedlejší větu vztažnou se vážou ke svému antecedentu v řídící větě. Vztažné vedlejší věty mají jednotné schéma: vztažný prostředek (zájmeno/zájmenné příslovce) odkazuje k substantivu, které je rozvíjeno danou vedlejší větou. Řídící sloveso vedlejší věty má funktor RSTR.

Výjimku tvoří konstrukce *způsob, jak...*, *cesta, jak...* a *možnost, jak...*: v těchto případech je funktor slovesa vztažné věty označen jako PAT.

4.2.2 Vztažná zájmena

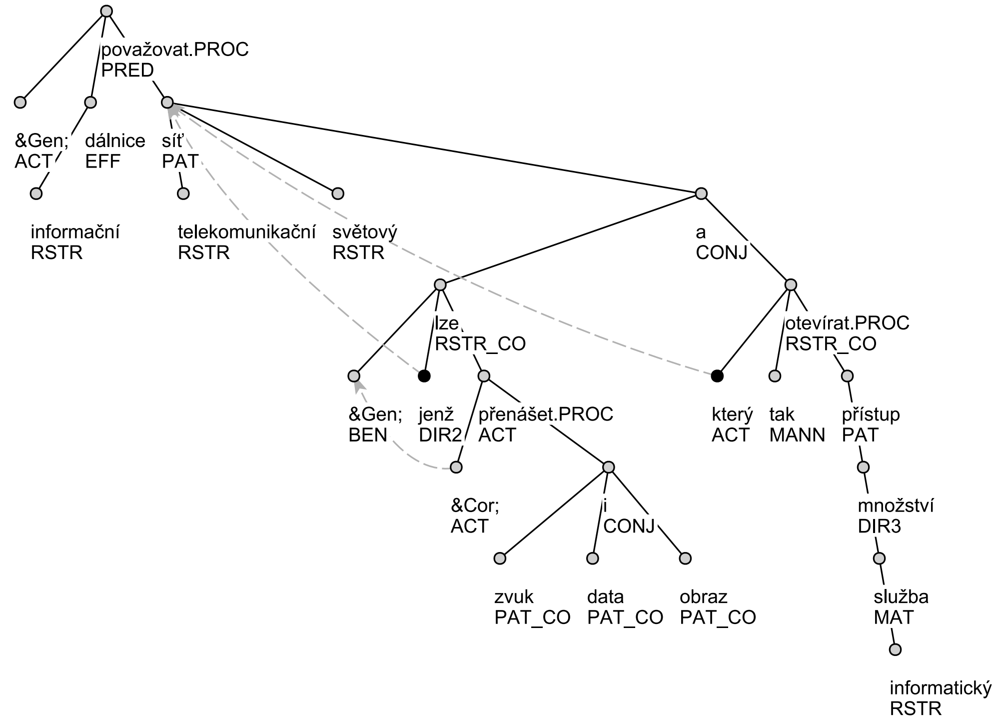
Koreferenci vztažných zájmen zachycujeme analogicky s příklady na obrázcích 4.6, 4.7 a 4.8.

4.2.3 Zájmenná příslovce (*kdy, kde, kam, jak, odkud*)

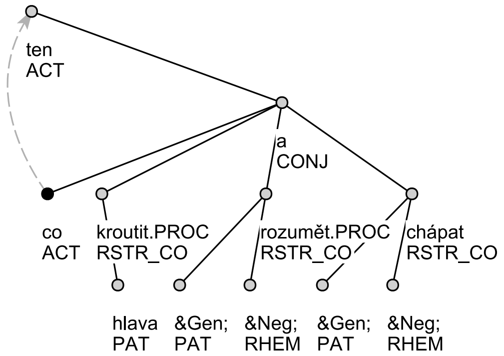
Koreferenci zájmenných příslovcí zachycujeme analogicky s příklady na obrázcích 4.9 a 4.10.

4.2.4 Nepravé vedlejší věty vztažné

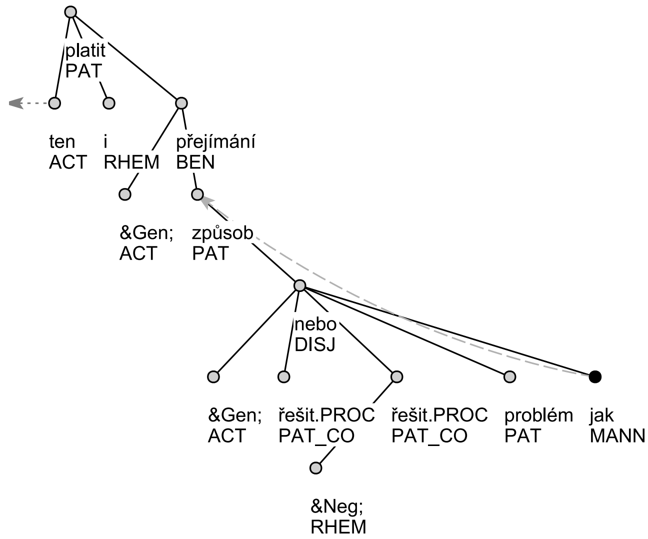
Koreferenční vztah zachycujeme i u nepravých vedlejších vět vztažných (vět pokračovacích), viz příklad na obrázku 4.11). Zachováváme záměr mluvčího a upřednostňujeme formální formu výpovědi (hypotaktické vyjádření koordinačního vztahu); sémantické posuny zanedbáváme. Anotátor koreference tuto informaci poznamená do atributu `corinfo`.



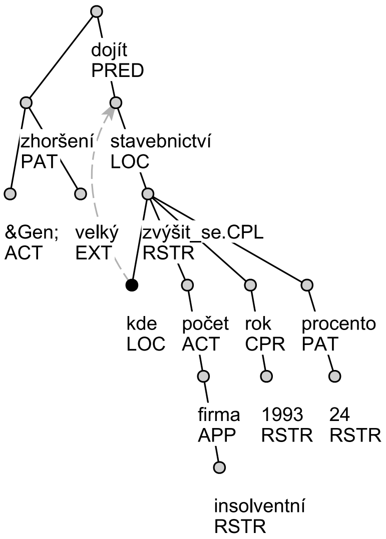
Obrázek 4.6: Vztažná zájmena: *Za informační dálnici se považuje světová telekomunikační síť, po **níž** lze přenášet zvuk, data i obraz a **která** tak otevírá přístup k množství informatických služeb.* (ln01#25)



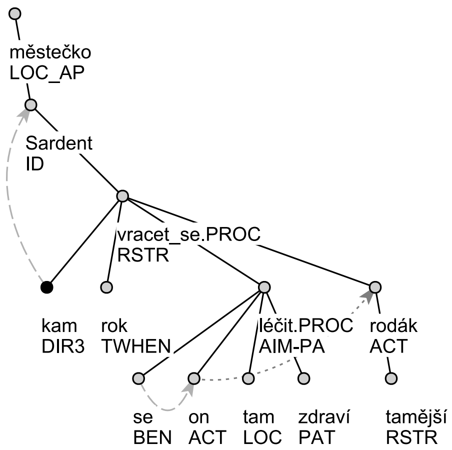
Obrázek 4.7: Vztažná zájmena: *Ti, **co** kroutí hlavami, nerozumí a nechápou, zároveň instinktivně varují.* (lo51#18)



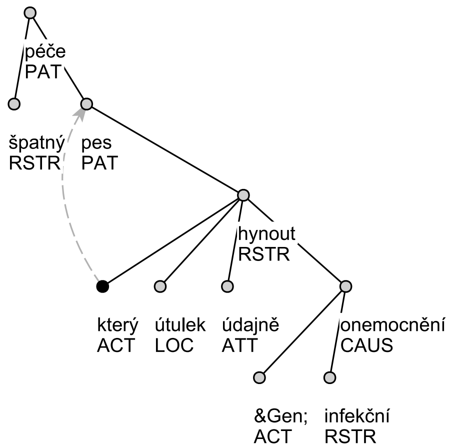
Obrázek 4.8: Vztažná zájmena: *Obávám se, že to platí i pro přejímání způsobů, **jak** řešit anebo neřešit problémy.* (lm37#15)



Obrázek 4.9: Zájmenná příslovce: *K největšímu zhoršení došlo v oblasti stavebnictví,* ***kde*** *se počet insolventních firem oproti roku 1993 zvýšil o 24 procent.* (ln01#44)



Obrázek 4.10: Zájmenná příslovce: *Film se odehrává na venkově, v městečku Sardent,* ***kam*** *se po letech vrací - aby si tam léčil zdraví - tamější rodák.* (lo51#48)



Obrázek 4.11: Nepravé vedlejší věty vztažné: *Představitelé Hnutí ochránců zvířat obvinili ředitelku Interpespenziónu ze špatné péče o psy,* ***kteří*** *v útulku údajně hynou na infekční onemocnění.* (ml42#20)

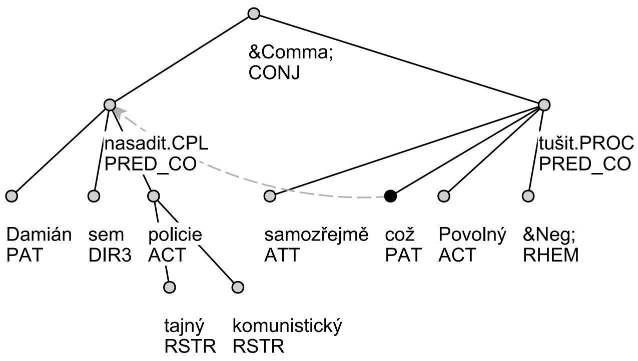
4.2.5 Spojovací výraz *což*

Koreferenční vztah zachycujeme v případech, kdy si *což* zachovává větněčlenskou platnost (i v nepřímých pádech, např. *bez čehož*, *čemuž* . . .).

Antecedentem je většinou **bezprostředně předcházející klauze**: koreferenční šipku vedeme k uzlu řídícího slovesa této klauze.

Rozlišujeme následující situace:

1. Z velké míry se jedná o stálou konstrukci: obě klauze jsou koordinačně spojeny a *což* odkazuje k levé sestře svého řídícího slovesa (obr. 4.12).
2. Další pravidelná konstrukce je rozvinutím té, kterou jsme uvedli jako převažující (viz obr. 4.12): zde je však koordinačně připojená věta sama rozdělena do několika (2) klauzí, jež jsou sémanticky v konfrontačním poměru (*zatímco, kdežto, ale*. . .). V těchto případech opět platí, že antecedentem je bezprostředně předcházející klauze (obr. 4.13).
3. Okrajově *což* neodkazuje k řídícímu slovesu bezprostředně předcházející klauze, ale k nějakému na něm závislému substantivu. Může jít jak o dějové substantivum (obr. 4.14), tak o běžné apelativum (obr. 4.15).



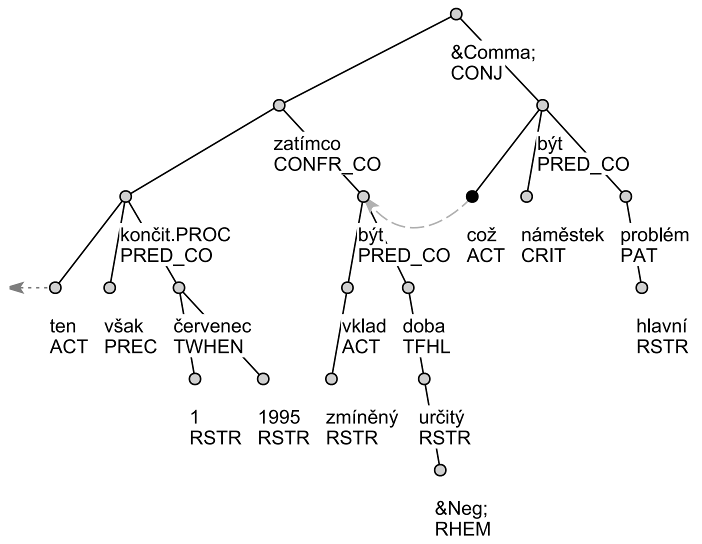
Obrázek 4.12: Spojovací výraz *což*: *Damiána sem nasadila komunistická tajná policie,* ***což*** *samozřejmě Povolný nemohl tušit.* (lm50#47)

Vzhledem k povaze spojovacího výrazu *což* se můžeme běžně setkat i s koreferencí, která přesahuje rámec graficky vymezené věty a funguje na úrovni nadvětné (textové). Protože jde v takových případech o čistě formálně rozdělené souvětí, můžeme i zde použít pravidlo směřování koreferenční šipky k předchozí klauzi: antecedentem bude řídící sloveso celé předchozí věty:[4]

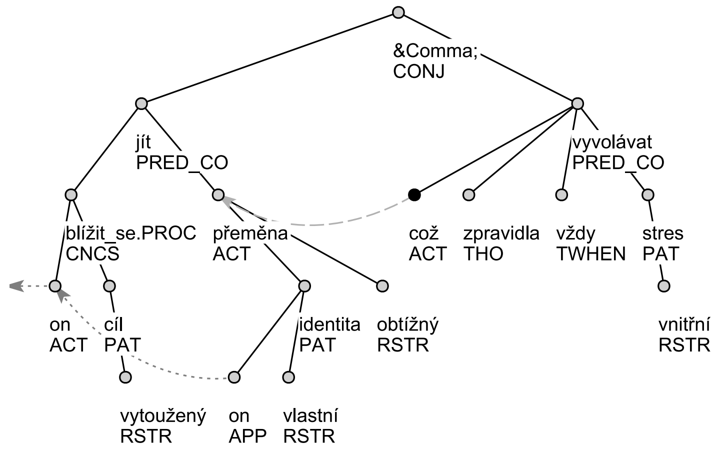
▷ ml8#5 *Právě vedoucí týmu Motorsport Škoda Pavel Janeba jen pokrčil rameny na otázku, jak dopadlo jednání uvnitř koncernu VW, které soutěže v příštím roce jeho tým absolvuje a které ne a kolik jich dohromady bude, jaká bude celková strategie.* ***Což*** *znamená, že vše je zatím ve hvězdách.*

▷ ml16#26 *Pan předseda Lux se nemůže smířit s tím, že podpora jeho křesťansky orientované strany je taková, jaká je.* ***Což*** *ho, myslím, vede ke křečovitým formulacím.*

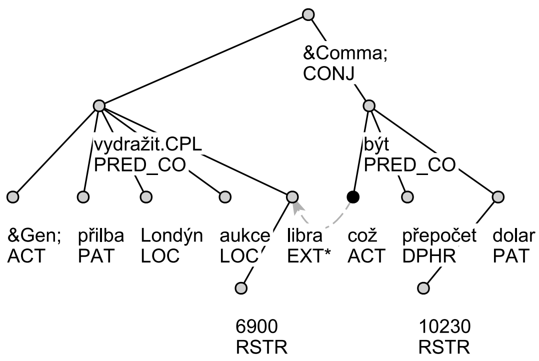
[4]Každý případ, kdy gramatická koreference překročí hranice věty, je silně příznakový, má velmi výraznou zdůrazňovací funkci. Z toho důvodu jsou takové výskyty spíše výjimečné, nicméně se s nimi v PDT setkáváme:
ll22#16 *Na dnešní den byl oznámen začátek soudního řízení s nejznámějším a nejvlivnějším politikem Itálie za posledních padesát let, sedminásobným premiérem a symbolem mocenského režimu Křesťanské demokracie. Mužem, na němž se Italům líbila obratnost, chytrost, intelekt, lehký smysl pro humor, schopnost kompromisu. A* ***kterého*** *ztotožňoval tu s érou hmotného vzestupu, tu s hnilobnou stranokracií.*



Obrázek 4.13: Spojovací výraz *což*: *Ty však končí k 1. červenci 1995, zatímco zmíněné vklady jsou na dobu neurčitou,* ***což*** *je podle náměstka hlavní problém.* (lo17#18)



Obrázek 4.14: Spojovací výraz *což*: *Ačkoli se blíží vytouženému cíli, jde o obtížnou přeměnu jeho vlastní identity,* ***což*** *zpravidla vždy vyvolává vnitřní stres.* (lm51#45)



Obrázek 4.15: Spojovací výraz *což*: *Přilba Nigela Mansella, loňského mistra světa formule 1 a vedoucího jezdce průběžného pořadí americké série IndyCar, byla vydražena v Londýně na aukci za 6900 liber,* ***což*** *je v přepočtu 10 230 dolarů.* (ml07#6)

4.3 Doplněk

Doplněk je takové doplnění slovesa, které se současně vztahuje i k podmětu, předmětu či jinému větnému členu (má dvojí sémantický vztah). Tyto dva typy závislosti jsou v tektogramatické stromové struktuře zachyceny různými prostředky:

- závislost na slovese je znázorněna hranou závislostního stromu (tedy stejně jako závislost jiných členů na slovese);
- závislost na „sémanticky řídícím" substantivu je reprezentována způsobem podobným zachycení koreferenčního vztahu;[5] je také znázorněna jako šipka „napříč" stromem.

Rozlišujeme následující typy doplňků:

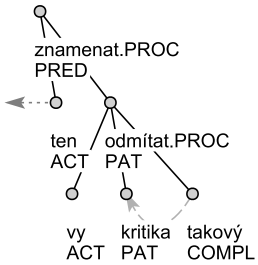
- doplněk vyjádřený neslovesnou formou (kapitola 4.3.1)
- doplněk vyjádřený slovesnou formou
  - doplněk vyjádřený neurčitým slovesným tvarem (kapitola 4.3.2): přechodník, participium, infinitiv
  - doplněk vyjádřený určitým slovesným tvarem (kapitola 4.3.3): vedlejší věta doplňková

4.3.1 Doplněk vyjádřený neslovesnou formou

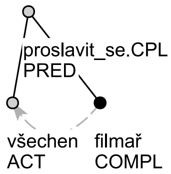
V případě vyjádření doplňku **neslovesnou formou** je druhá sémantická závislost na substantivu v jazyce vyjádřena vždy také formálně: adjektiva se shodují v rodě, v čísle i v pádě, substantivum ve funkci doplňku se se svým „řídícím" substantivem shoduje v rodě a v čísle pouze selektivně. Pád substantiva je dán u prostých doplňků vazbou slovesa (má lexikalizovanou formu – typ *Seděla mu modelem*), u doplňků se spojkou *jako* (*jakožto*, *coby*) se shoduje s pádem „sémanticky řídícího" substantiva.

U doplňku vyjádřeného **substantivem, adjektivem** či **číslovkou** vede šipka přímo od uzlu daného doplňku (obrázky 4.16, 4.17 a 4.18).

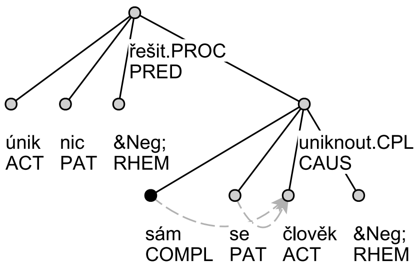
[5]Jde jen o technické řešení, ve skutečnosti se nejedná o koreferenční vztah.



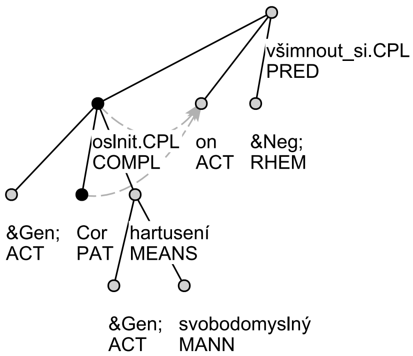
Obrázek 4.16: Doplněk vyjádřený neslovesnou formou: *Znamená to, že odmítáte kritiku jako **takovou**?* (ml05#7)



Obrázek 4.17: Doplněk vyjádřený neslovesnou formou: *Všichni se proslavili jako **filmaři**.* (lo52#16)



Obrázek 4.18: Doplněk vyjádřený neslovesnou formou: *Únik nic neřeší, protože **sám** před sebou člověk uniknout nemůže.* (lm52#1)



Obrázek 4.19: Doplněk vyjádřený participiem: ***Oslněn** svobodomyslným hartusením nevšimne si, že dvě stě nezávislých kandidátů nepředstavuje nic jiného než dvě stě politických stran.* (lo24#17)

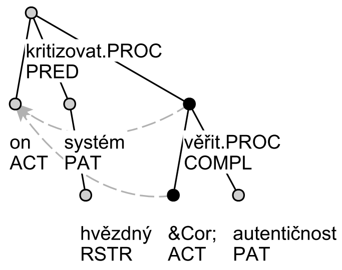
4.3.2 Doplněk vyjádřený neurčitým slovesným tvarem

Je-li doplněk vyjádřen **neurčitým slovesným tvarem**, hovoříme nejen o závislosti, ale i o referenci. Sémanticky řídící substantivum je totiž zároveň ve vztahu gramatické koreference s jedním z valenčních členů slovesa, které má funktor COMPL. Tento valenční člen má proto vždy lemma *Cor* (viz kapitola 4.4) a vede od něj koreferenční šipka.[6]

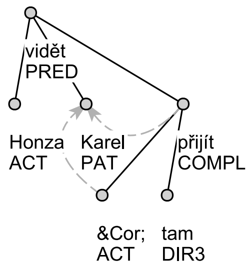
U doplňků vyjádřených **participiem** může být ve vztahu gramatické koreference jak Aktor participia, tak i Patiens, případně jiný valenční člen (obr. 4.19).

U doplňků vyjádřených aktivním **přechodníkem** je ve vztahu gramatické koreference vždy Aktor přechodníku a subjekt řídícího slovesa (obr. 4.20), u doplňků vyjádřených pasivním přechodníkem *Odcházel, byv poražen*) může být ve vztahu gramatické koreference i Patiens přechodníku (případně jiný aktant) a subjekt řídícího slovesa.

U doplňků vyjádřených **infinitivem** (slovanský akuzativ s infinitivem) je antecedentem akuzativní objekt řídícího slovesa (obr. 4.21).



Obrázek 4.20: Doplněk vyjádřený přechodníkem: *Kritizovali hvězdný systém, **věříce** v autentičnost dosud neokoukaných tváří, které se však záhy také staly hvězdami.* (lo52#20)



Obrázek 4.21: Slovanský akuzativ s infinitivem: *Honza viděl Karla **přijít**.* (0)

4.3.3 Doplněk vyjádřený určitým slovesným tvarem (vedlejší věta doplňková)

Je-li doplněk vyjádřen vedlejší větou připojenou vztažným příslovcem ***jak***, je antecedentem Patiens řídícího slovesa, viz obrázek 4.22. Jinak postupujeme u vedlejších vět doplňkových připojených vztažným zájmenem ***jaký*** (považujeme za ně jen ty věty připojené zájmenem *jaký*, které stojí v typické pozici adjektivního doplňku; srov. následující dvě věty: *Přidělili nám vedoucího, jaký se jim namanul.COMPL* × *Vedoucí, jakého nám přidělili.RSTR, nestojí za nic.*). Zde je valenční rámec slovesa obsazen buď lexikálně, nebo je pozice pro Aktor vyplněna zájmenným lemmatem *on* (např. obr. 4.23) – pak postupujeme podle pravidel textové koreference.

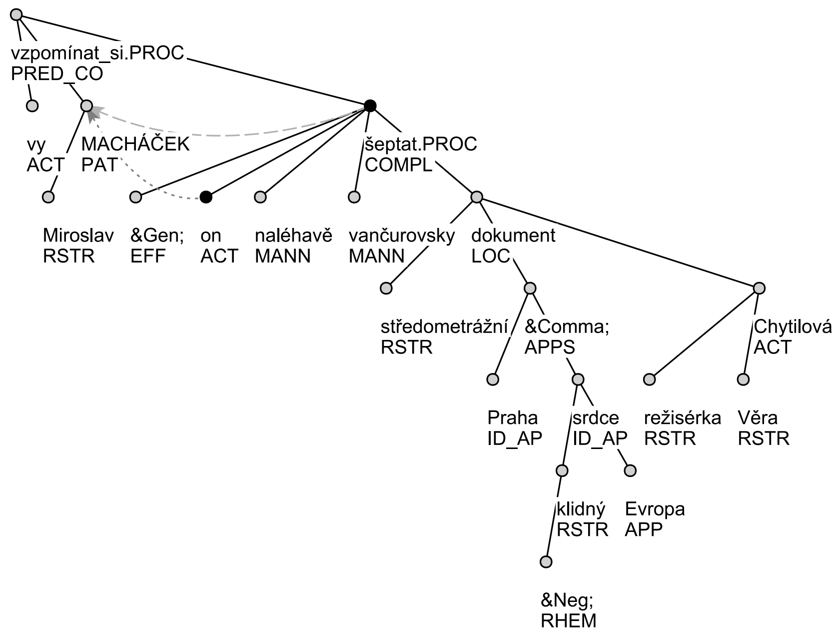
Stejně jako k infinitním slovesným tvarům přistupujeme i k **elidovaným přechodníkovým konstrukcím**. Uzel pro prázdné sloveso s funktorem COMPL (Emp.COMPL, viz obr. 4.24) může nahrazovat přechodník „*řka*“ u přímé řeči (*Vtrhl do dveří: „Kdy bude večeře?“*[7]) nebo přechodník „*maje*“ (např. po slovesech *stát*, *ležet*, *jít*, *mluvit*, *hledět*, *zpívat* aj.: *Jana seděla hlavu skloněnou*).

U ustrnulých, nekongruentních přechodníků (např. *takříkajíc*, *ne/chtě*, *stoje*, *leže*, *kleče* apod.) o koreferenci neuvažujeme. Považujeme je za adverbia, která si ponechala omezenou část slovesného valenčního potenciálu.[8]

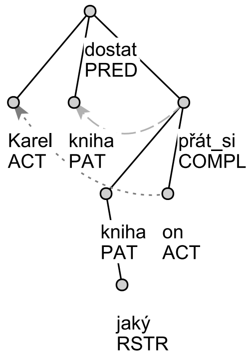
[6] Protože jmenné kategorie přechodníku a participia (rod a číslo) se shodují s antecedentem, bude jej možné určit automatickou procedurou. Nejčastěji jím bývá Aktor nebo Patiens.

[7] Srov. dodatek o přímé řeči v Manuálu pro tektogramatické značkování (ve všech dalších odkazech bude tato publikace pro zjednodušení označena jako manuál).

[8] Srov. dodatek o ustrnulých přechodnících v manuálu.



Obrázek 4.22: Doplněk vyjádřený vedlejší větou: *Vzpomínáte si na Miroslava Macháčka, jak naléhavě vančurovsky **šeptal** ve středometrážním dokumentu Praha, neklidné srdce Evropy režisérky Věry Chytilové?* (ls23#37)



Obrázek 4.23: Doplněk vyjádřený vedlejší větou: *Karel dostal knihu, jakou si **přál**.* (0)

4.4 Kontrola

Kontrola je vztah obligatorní nebo fakultativní referenční závislosti mezi kontrolujícím členem (termín *Controller*) a členem kontrolovaným (*Controllee*).[9]

Controller je jedním z členů valenčního rámce řídícího slovesa: např. u slovesa *plánovat* plní roli ACT, u slovesa *radit* má roli ADDR, u slovesa *poslat* PAT. V některých konstrukcích se setkáme i s aktantem BEN (např. *je nutné*). Výjimečně může být Controllerem i volné určení místa, např. LOC (*Být dobře zapsán u šéfa v něm.LOC vyvolávalo pocit hrdosti*).

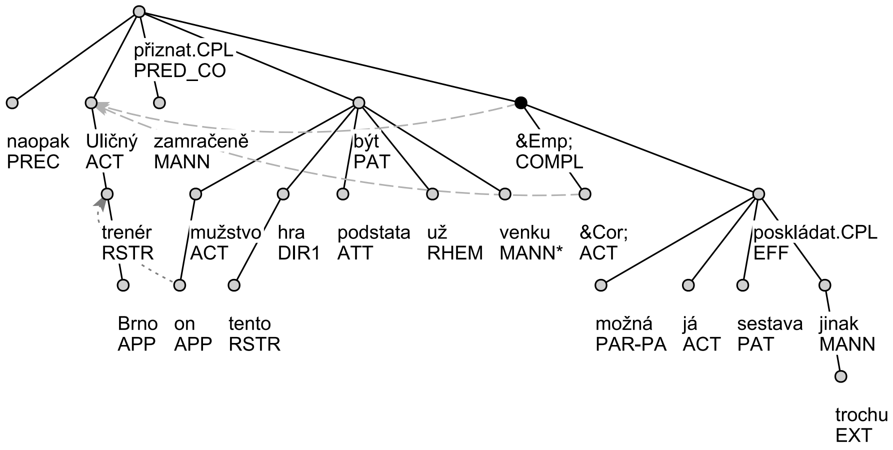
Controllee je vždy nevyjádřený subjekt infinitivu nebo slovesného derivátu závislého na řídícím slovese; *Controllee* má nejčastěji funktor ACT, může však mít i funktor PAT a ADDR (např. u pasivního infinitivu nebo u infinitivu zastupujícího pasivní konstrukci). *Controllee* přiřazujeme lemma *Cor*, viz obrázky 4.25 a 4.26.

Jak řídící sloveso, tak infinitiv se mohou nominalizovat. Možné typy konstrukcí s kontrolou podrobněji viz kapitola 4.4.1.

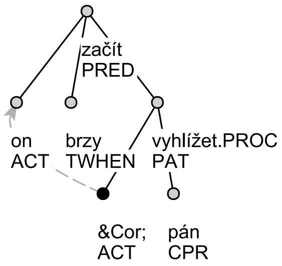
Vztah kontroly je podmíněn lexikálním významem řídícího slovesa, předpokládáme tedy, že je potenciálně možné sestavit seznam sloves kontroly. Přesto není v některých případech možné určit kontrolu automaticky: roli zde hrají nejen lexikální záležitosti jako polysémie, metonymie apod. (viz níže), ale i konkrétní použití slovesa. Někdy je tedy třeba řídit se kontextem.

Úkolem anotátora koreference je zachytit koreferenční vztah mezi *Controllee* a *Controllerem*, a to i v případech, kdy anotátor velkého souboru nedoplnil uzel pro *Controllee* (např. u substantiv a adjektiv). Protože anotátor koreference pracuje více s kontextem, může o ne/zachycení koreferenčního vztahu rozhodnout v rozporu s anotací

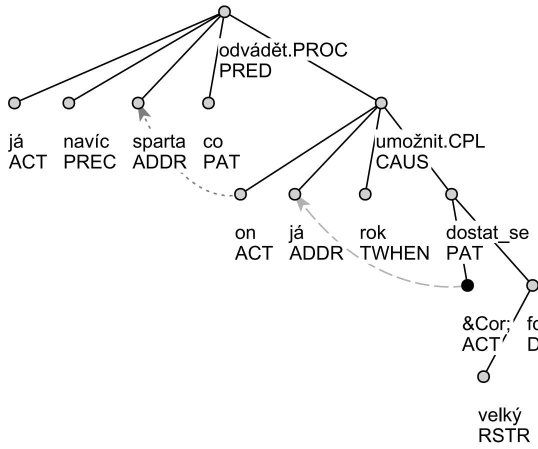
[9]Tyto termíny použil poprvé Chomsky v Theory of Binding. Jeho příliš úzké pojetí zanedbáváme a pracujeme s přístupem používaným v rámci Funkčního generativního popisu, viz [Koktová, 1992] a [Panevová, 1996].



Obrázek 4.24: Elidovaný přechodník: *Trenér Brna Uličný naopak zamračeně přiznal, že jeho mužstvo je z této hry už v podstatě venku: „Možná, že jsem měl sestavu poskládat trochu jinak.“* (lm46#49)



Obrázek 4.25: Kontrola: *Pokud dámy postupují podobně, **začnou** brzy **vyhlížet** jako pánové.* (lm37#31)



Obrázek 4.26: Kontrola: *Navíc mám Spartě co odvádět za to, že mi před lety **umožnila dostat** se do velkého fotbalu.* (ml09#14)

velkého souboru – v tom případě je v jeho kompetenci změnit lemma uzlu (např. Gen → on/Unsp/Cor a opačně). Jedná se zejména o případy, kdy je nutné se rozhodnout, zda jde o konstrukci s kontrolou, nebo o textovou koreferenci, případně o vyjádření se všeobecným aktantem, a to jak u infinitivu, tak (zejména) u vágnějších nominalizací. Podobně postupuje u polysémních výrazů (na základě kontextu volí konkrétní význam lexie, např. *chtít*$_1$*=toužit [chce pracovat], chtít*$_2$*=vyžadovat [Anežka chce číst pohádky]*) a „metonymických“ vyjádření (viz níže). Případnou nejistotu o tom, kdo je skutečným konatelem, poznamená anotátor koreference do atributu `corinfo`.

4.4.1 Typy konstrukcí s kontrolou

Ve většině případů platí, že jak řídící sloveso, tak infinitiv se mohou nominalizovat. Existují tedy následující typy konstrukcí s kontrolou:

1. infinitiv závislý na slovesném predikátu kontroly
   (a) infinitiv závislý na syntetickém (jednoslovném) predikátu kontroly
   (b) infinitiv závislý na složeném predikátu kontroly
2. infinitiv závislý na nominalizaci slovesného predikátu kontroly (tj. na deverbativním substantivu nebo adjektivu)
3. nominalizace infinitivu, příp. nominalizace vedlejší věty závislá na slovesném predikátu kontroly
4. nominalizace infinitivu, příp. nominalizace vedlejší věty závislá na nominalizaci slovesného predikátu kontroly.

Většina sloves kontroly se může vyskytnout ve všech čtyřech typech výše uvedených konstrukcí (tedy např.: *slíbit napsat dopis, slib napsat dopis, slíbit napsání dopisu, slib napsání dopisu*). Některá slovesa kontroly (např. *přisoudit, osočit, podezírat, stíhat*) však vůbec nemohou být rozvita infinitivem, tudíž se mohou vyskytnout pouze v konstrukcích typu 3 a 4 (např. *podezírat z krádeže, podezření z krádeže*, ale **podezírat krást, *podezření krást*). Ve výjimečných případech není možná ani nominalizace infinitivu, ani nominalizace řídícího slovesa (např. *Viktor se zdá být chytrý*), takové sloveso kontroly se tedy může vyskytnout pouze v 1. skupině.[10]

4.4.2 Jednotlivé problematické okruhy konstrukcí s kontrolou vzhledem ke koreferenci

Problematika nominalizací v konstrukcích s kontrolou

Protože je substantivum ze své povahy výrazně vágnější než infinitiv, doprovází rozhodování, zda je nějaké konkrétní spojení slovesa se substantivem (případně substantiva se substantivem) skutečně konstrukcí s kontrolou, několik problémů:

- Ve spojeních slovesa se substantivem nemusí být zřejmé, zda tato konstrukce byla opravdu odvozena od spojení slovesa s infinitivem, nebo zda byla odvozena od spojení slovesa s nějakou vedlejší větou. Např. věta *Mikolášek se vyhýbá jednoduchému ztvárnění svých nápadů* může mít dvě interpretace: (1) *Mikolášek se vyhýbá jednoduše ztvárnit své nápady* – jde o kontrolu, (2) *Mikolášek se vyhýbá tomu, aby někdo ztvárnil jeho nápady jednoduše* – nejde o kontrolu;
- Substantivum může zanedbávat reflexivitu svého základového slovesa, např.: konstrukce *jeho rozhodnutí zrušit výrobu* může být odvozena od dvou konstrukcí: (1) *rozhodl se zrušit* – jde o kontrolu; (2) *rozhodl o zrušení výroby*, tj. *rozhodl, že někdo má zrušit výrobu* – nejde o kontrolu; v konstrukci *stíhají ho pro nedovolené ozbrojování* jde o kontrolu, ale není jasné, který z valenčních členů substantiva *ozbrojování* má mít lemma *Cor*, protože nevíme, zda základová konstrukce byla *ozbrojovat někoho*, nebo *ozbrojovat se.*[11]

[10] Více viz [Panevová et al., 2002].

[11] V těchto případech platí, že pokud nelze jednoznačně rozhodnout ani na základě kontextu, upřednostňujeme předpoklad, že substantivum bylo odvozeno od nereflexivního slovesa.

„Slovesa záměru“ a „slovesa přebírání zodpovědnosti“

Slovesa, složené predikáty, případně substantiva „záměru“[12] chápeme jako slovesa kontroly, přestože konstrukce s těmito slovesy můžeme někdy interpretovat i tak, že ten, kdo „má záměr“ něco udělat, nemusí nakonec skutečně danou činnost provádět. Za danou činnost má však „zodpovědnost“. Např. *Vedení sekce plánuje (má v plánu / má plán) vyklidit knihovnu; Pan Moric si vytkl za cíl proniknout na neobsazené trhy přijatelné pro Radu bezpečnosti* (z kontextu víme, že pronikat budou zbrojovky, nejen jejich ředitel Moric, ale přesto na sloveso *proniknout* zavěsíme uzel s lemmatem *Cor*, od nějž povede koreferenční šipka k lemmatu *pan*). Možnou rozdílnost konatelů zanedbáváme s tím, že to, kdo skutečně danou činnost provede, je plně v kompetenci toho, kdo „má záměr“ něco udělat.

Obdobně jako u sloves záměru postupujeme i u „metonymických“ užití některých jiných „sloves přebírání zodpovědnosti“; totožnost konatelů řídícího slovesa a závislého infinitivu nemusí být zcela zřejmá, přesto tuto konstrukci rozebíráme jako konstrukci s kontrolou (např. *slíbil zapůjčit* může být interpretováno také jako *slíbil zajistit, že někdo jiný zapůjčí. . .*). Případnou nejistotu o tom, kdo je skutečným konatelem, poznamená anotátor koreference do atributu `corinfo`.

Složené predikáty kontroly

Za složené predikáty kontroly považujeme všechna synonymní víceslovná vyjádření sloves kontroly.

- Jde zvláště o **kvazimodální slovesa**,[13] tj. synonymní vyjádření modálních sloves (např. *mít schopnost, chuť, dar, potřebu, šanci, příležitost; být schopen, ochoten, povinen*) a sloves vyjadřujících záměr (např. *mít v úmyslu (úmysl), záměr; mít v plánu (plán); mít tendenci; být připraven, odhodlán*);
- Do této kategorie řadíme i **kvazifázová slovesa**[14] (např. *sbírat odvahu, dostat chuť*) a slovesa s významem *„umožnit někomu udělat něco“*, a to jak v „aktivním“ užití (tj. *on umožnil. . .*), tak v „pasivním“ užití (tj. *bylo mu umožněno*), tedy např. *dát někomu šanci (příležitost) udělat něco*, ale i *dostat od někoho šanci (příležitost) udělat něco*;
- Za zvláštní typ složených predikátů považujeme i některé **verbonominální predikáty** se sponovým slovesem *být* (např. *Jeho úkolem je nalézt řešení*; *Petr je ochoten přijít*), srov. níže.

U složených predikátů tvořených slovesem a substantivem[15] dochází k referenční totožnosti určitých valenčních členů daného substantiva a slovesa.[16] Jde nejčastěji o totožnost Aktora substantiva s Aktorem řídícího slovesa, ale např. u konstrukcí, v nichž je daný složený predikát synonymním vyjádřením pasivní konstrukce, jde o totožnost Aktora substantiva s Origem slovesa (např. *dostat od někoho.*ORIG *příkaz*). U sloves, která si i po svém významovém vyprázdnění uchovávají dativní valenci, může při jejich rozvití deverbativním substantivem, které má rovněž dativní valenci, dojít jak k totožnosti Aktora substantiva s Aktorem slovesa, tak k totožnosti Adresátu substantiva s Adresátem slovesa (např. *poskytnout radu/službu/půjčku; dát radu/příkaz*). Relevantní totožné uzly u substantiva mají speciální lemma *QCor* (tj. *Quasi-Control*, obr. 4.27).[17]

---

[12]Pojmy „slovesa záměru“ a „slovesa přebírání zodpovědnosti“ zavádíme pouze jako pracovní termíny.

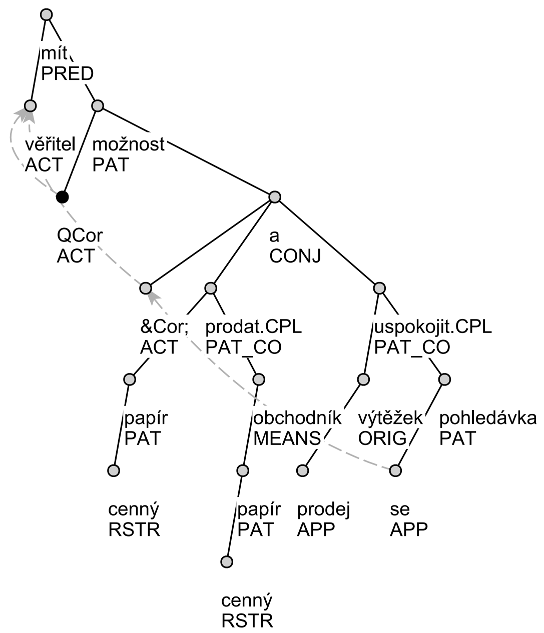
[13]Viz manuál.

[14]Pracovní termín.

[15]Jmenné části (substantivu) složeného predikátu přiřazujeme speciální funktor CPHR (tj. *compound phraseme*).

[16]Srov. dodatek o složených predikátech v manuálu.

[17]Není-li dané spojení slovesa a substantiva považováno za složený predikát, příp. jiný druh frazému, má tato jmenná část funktor, který by měla při bezpříznakovém užití slovesa. Jmenné části je přiřazen valenční rámec odpovídající užití daného substantiva a ten její valenční člen, který by v konstrukcích se složenými predikáty dostal lemma *QCor*, dostane lemma odpovídající danému koreferenčnímu vztahu (*on*/*Gen*/*Unsp*).



Obrázek 4.27: Složené predikáty kontroly: *Věřitel* ***má možnost*** *cenný papír* ***prodat*** *prostřednictvím obchodníka s cennými papíry a z výtěžku prodeje* ***uspokojit*** *svou pohledávku.* (lo08#2)

Vzhledem k tomu, že složený predikát je jedním z typů frazémů, tj. chová se částečně jako jedna lexikální jednotka, platí při zachycování koreferenčních vztahů v těchto konstrukcích následující zásady:

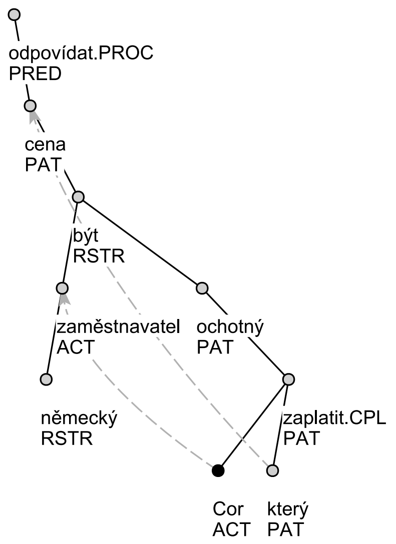
- K uzlu s lemmatem *QCor* nevede žádná koreferenční šipka;
- U uzlu s lemmatem *QCor* je zaznačen koreferenční vztah k uzlu, se kterým je uzel s lemmatem *QCor* totožný;
- U uzlu pro *Controllee* (s lemmatem *Cor*) je zaznačen koreferenční vztah k uzlu pro *Controller*.

Infinitiv závislý na „složeném predikátu kontroly" tvořeném slovesem a adjektivem

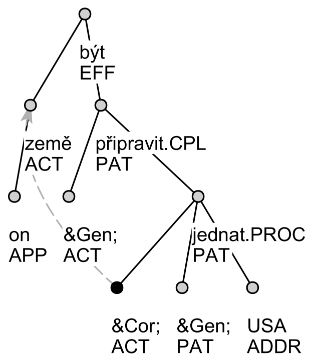
Tento typ složených predikátů kontroly reprezentují zejména konstrukce tvořené sponovým slovesem *být* a predikativními adjektivy, např. *schopný, ochotný, povinen, povinován, vědomý_si, zvyklý, náchylný, připravený, rozhodnutý, způsobilý* (obr. 4.28 a 4.29). Infinitiv zde závisí na adjektivu. Adjektivum (jmenná část složeného predikátu) nemá žádný speciální funktor (zpravidla dostane funktor PAT). *Controllerem* je tedy v těchto konstrukcích vždy ACT slovesa *být*.

Infinitiv závislý na slovesné části „složeného predikátu kontroly" jako ACT

V konstrukcích, kde infinitiv závisí na slovesné části „složeného predikátu kontroly" jako ACT, jde o kontrolu fakultativní. Specifikem těchto konstrukcí je to, že kontrolujícím členem (antecedentem) je zde nejčastěji Benefaktor (BEN). Členem kontrolovaným je jeden z členů valenčního rámce infinitivního tvaru slovesa, nejčastěji ACT.



Obrázek 4.28: Složené predikáty kontroly: *Čisté měsíční výdělky okolo 850 až 1400 marek odpovídají ceně, kterou* ***jsou*** *němečtí zaměstnavatelé* ***ochotní zaplatit***. (lo10#34)



Obrázek 4.29: Složené predikáty kontroly: *Reagoval na prohlášení kubánského ministra zahraničí Roberta Robainy, který řekl, že jeho země* ***je připravena jednat*** *s USA. . .* (lo15#35)

„Složeným predikátem kontroly“ zde chápeme celou škálu **verbonominálních predikátů**, jejichž slovesnou část tvoří sponové sloveso *být* (také viz obr. 4.30 a 4.31):

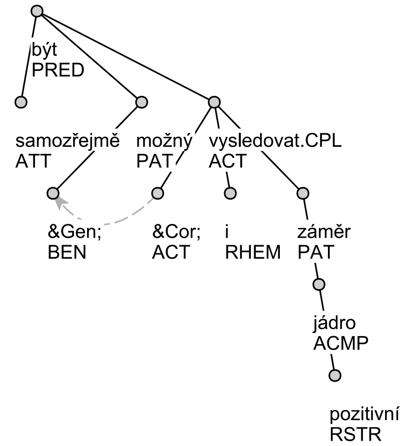
- sloveso + adjektivum modální (např. *být možné udělat*)
- sloveso + adjektivum hodnotící (*být jednoduché udělat*)
- sloveso + substantivum hodnotící (např. *být radost udělat*)
- sloveso + adjektivum „osobního prožívání“ (*být trapné udělat*)
- sloveso + adverbium „osobního prožívání“ (např. *být zatěžko udělat*)
- sloveso + substantivum ze složených predikátů (např. *být povinnost(í)/úkol(em) udělat*)
- sloveso + predikativní adverbia (např. *být možno, nutno, třeba*)

Do této kategorie dále řadíme některé frazémy (např. *Nestálo za úvahu.*DPHR *půjčit.*ACT *si na posily od banky*), konstrukce tvořené slovesem *být* + infinitivem sloves smyslového vnímání a poznávání + akuzativem jména (tj. typ *Je vidět Sněžku*[18]) a sloveso *lze* (viz obr. 4.32).

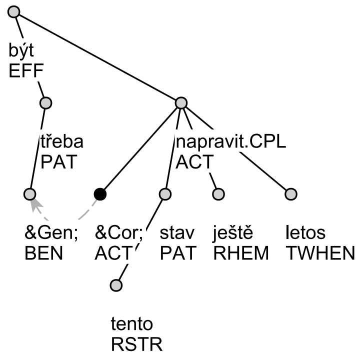
4.4.3 Infinitivní konstrukce, v nichž nejde o kontrolu (přesahy do textové koreference)

Subjekt infinitivu zde vstupuje do koreferenčních vztahů, ale nejde o vztah gramatické koreference spojený s kontrolou. Anotátor koreference rozhodne podle kontextu o tom, zda se jedná o textovou koreferenci (pak uzlu s funktorem ACT pod infinitivem přiřadí lemma *on* a zachytí koreferenční vztah) nebo zda je na místě vyjádření s Gen.ACT. Rozlišujeme dva případy:

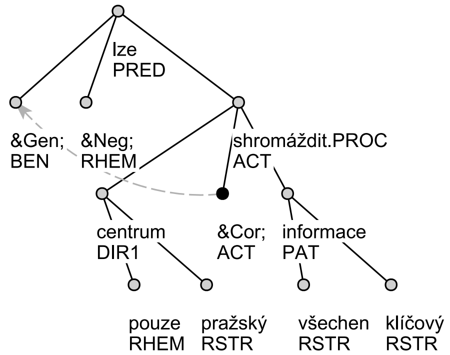
[18]Viz příslušný dodatek v manuálu.



Obrázek 4.30: Složené predikáty kontroly: *Samozřejmě* ***je možno vysledovat*** *i záměry s pozitivním jádrem.* (lm55#16)



Obrázek 4.31: Složené predikáty kontroly: *Domnívám se, že* ***je třeba*** *tento stav* ***napravit*** *ještě letos, říká.* (lo07#23)



Obrázek 4.32: *Pouze z pražského centra* ***nelze shromáždit*** *všechny pro investory klíčové informace.* (lo06#22)

1. **Infinitiv závisí na slovese, které není slovesem kontroly.** Nevyjádřený subjekt infinitivu nebo jeho nominalizace bude mít lemma odpovídající danému koreferenčnímu vztahu.
   - ▹ *Rozhodl zrušit výrobu.*
   - ▹ *Zakotvit do ústavy trvale vyrovnaný rozpočet nepovažuje za nejšťastnější místopředseda sněmovny Jiří Vlach.*
   - ▹ *Proto považujeme za klíčovou otázku tento systém změnit.*

2. **Konstrukce se slovesy umožňujícími dvě infinitivní doplnění.** Do této skupiny patří zejména slovesa *být* a *znamenat*. Nevyjádřený subjekt infinitivu (příp. jeho nominalizace) bude mít lemma odpovídající danému koreferenčnímu vztahu.
   - ▹ *Napsat článek pro mě znamená měsíc nedělat nic jiného.*
   - ▹ *Ustupovat jim znamená vracet se ke státem řízené ekonomice.*
   - ▹ *Dělat to takto by bylo nošením dříví do lesa.*

Kapitola 5

Textová koreference

Textová koreference je běžně chápána jako užití různých jazykových prostředků (zájmena, synonyma, zobecňující substantiva aj.), které anaforicky odkazují. Toto odkazování není dáno gramaticky, ale na základě kontextu. Prostředky textové koreference jsou svou povahou vágní a určení antecedentu pouze na základě kontextu je problematické, proto se v našem pojetí soustředíme prozatím pouze na nejčastější prostředky textové koreference, tedy na **zájmena**. Pracujeme se zájmeny osobními a ukazovacími (zájmeno *ten*), která poukazují primárně k vnitrotextovým objektům; zájmena 1. a 2. osoby ponecháváme stranou.

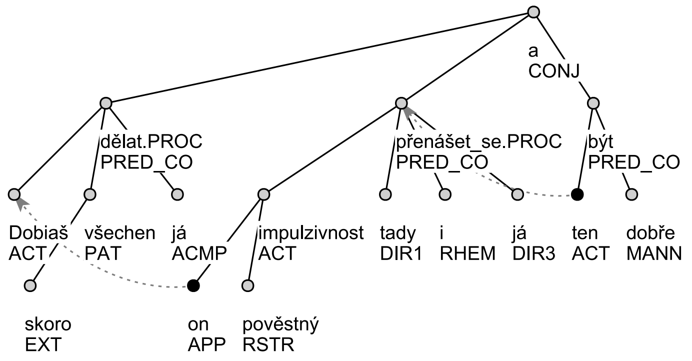
U zájmena *ten* nebereme v úvahu jeho atributivní použití.

5.1 Typy textové koreference

5.1.1 Explicitní antecedent

O explicitním antecedentu hovoříme v případech, kdy lze jako antecedent zájmena označit konkrétní podstrom tektogramatické struktury dané věty nebo některé z okolních vět. Jak bylo uvedeno v kapitole 3.1.2, koreferenční šipku vedeme pouze ke kořenu tohoto podstromu.

Na obrázku 5.1 je věta, ve které je antecedentem zájmena *jeho* uzel *Dobiaš* (triviální podstrom s jediným uzlem bez potomků), zatímco antecedentem zájmena *to* je celá klauze *jeho pověstná impulzivnost se přenáší i na nás*.



Obrázek 5.1: *Dobiaš skoro všechno dělá s námi,* ***jeho*** *pověstná impulzivnost se přenáší i na nás, a* ***to*** *je dobře.* (ml19#21)

5.1.2 Segment (Segm)

Odkazuje k většímu úseku textu. Jedná se o tyto případy:

1. Antecedentem je dvě a více konkrétních vět. Jde zejména o prostý výčet jednotlivostí, soubor za sebou jdoucích vět. Neodkazujeme zvlášť ke každé z nich, ale ke všem dohromady jako celku:
   - ▷ *Rozprava o podobě reformy veřejných financí bude zahájena ve středu. Všechna jednání proběhnou za zavřenými dveřmi. Lidovým novinám* ***to*** *sdělil včera ministr financí.*
   - ▷ *Podle Kohla nelze zapomenout na to, že Německo přepadlo 22. června 1941 Sovětský svaz. Němci jménem Německa přivodili ruskému lidu nesmírné utrpení. Stejně tak nelze zapomenout, co Rusové později způsobili Němcům. Z* ***toho*** *všeho si chceme vzít společné poučení.*
2. Antecedent není možno určit jako konkrétní uzel, ale lze jej inferencí vyvodit z úseku textu na základě kontextu:
   - ▷ *Předsedové a ekonomové družstev už jsou nachystaní na likvidaci dlužníků. Řekněme, že přijdou za vlastníkem 25 ha v družstvu. Každý ten hektar má hodnotu okolo 100 tisíc. Banka nabídne 10 tisíc za hektar a vlastníkovi nezbude nic jiného, než to prodat, protože nazítří mu banka nabídne už třeba jen 8 tisíc. Chci* ***tím*** *říci, že nebude všechno v transformačním procesu úplně čisté. Potentáti v bance koupí za deset, prodají si za patnáct. Ale povede to k rychlému přerodu. Zmizí výměry kolem 25 ha, přibude vlastníků kolem 500. Odhaduji, že do dvou let budou schopni splatit bance dluh a třetím rokem už budou dělat na sebe. A na práci najmou jen schopné lidi, bude to v jejich zájmu. Kdo* ***to*** *pochopil, má náskok.*

5.1.3 Exofora (Exoph)

U exofory jde o deiktické odkazování: zájmeno poukazuje k mimotextové situaci či skutečnostem:

- ▷ *V období vrcholícího léta roku 1939 již málokdo v Evropě mohl uvěřit nadějeplným slovům britského ministerského předsedy Chamberlaina, proneseným z balkonu Buckinghamského paláce po návratu z Mnichova: Myslím, že je* ***to*** *mír na celou naši dobu.*
  (*ten* = Mnichovská dohoda)
- ▷ ***Obětování [on.ACT] [on.BEN] Izáka.*** *Příběh přeživšího. Je to matoucí příběh, v němž vládne strach. Strach a víra. Strach a výzva. Strach a smích. V celé své hrůze se tento příběh stal zdrojem útěchy pro všechny, kteří ho přijali za svůj a předávají ho dále, vtiskujíce do něj svou vlastní zkušenost.*
  (*on*.ACT = Abrahám, *on*.BEN = Bůh)

5.2 Lemmata z pohledu koreference

S textovou koreferencí souvisejí nejen trlemmata *ten* a *on*, u nichž koreferenci primárně předpokládáme. Svou roli hrají také lemmata doplněných (tj. povrchově nevyjádřených) uzlů.[1] Pro odlišení doplněných uzlů, které mají explicitní antecedent, uzlů s „obecným“ antecedentem a uzlů, u nichž antecedent (prozatím) nepředpokládáme, je velmi důležitá také distinkce mezi typy elips, pro něž používáme lemmata *Gen* a *Unsp*.[2]

[1] Viz [Panevová, 1998].
[2] Pro podrobnější pohled na danou problematiku srov. [Řezníčková, 2002].

5.2.1 Gen(eral)

Lemma *Gen* využíváme pro označení tzv. všeobecných aktantů,[3] a to jak u sloves, tak u substantiv, případně adjektiv. Typickou povrchovou realizací pro vyjádření všeobecného Aktora je reflexivní pasivum. Vzhledem k tomu, že referentem všeobecných aktantů je skupina všech lidí (příp. objektů) typických pro danou situaci, antecedent není možné v textu najít, a koreferenci tedy u tohoto typu elipsy nezachycujeme.

- ▷ *Tato potvrzení se vydávají* ***[Gen.ACT] [Gen.ADDR]*** *na počkání.*
- ▷ *Náš chlapec už čte* ***[Gen.EFF]****.*
- ▷ *Tento tuk je vhodný na pečení* ***[Gen.ACT] [Gen.PAT]****.*

5.2.2 Unsp(ecified)

Lemma *Unsp* tvoří při anaforickém odkazování přechodovou fázi mezi nekoreferujícím výrazem a textovou koreferencí.[4] Používá se v případech, kdy antecedent doplněného uzlu[5] nelze přesně určit: odkazuje spíše ke kontextu předchozího textu než ke konkrétní jednotce,[6] proto nevolíme lemma *on*, které má jasný anaforický charakter. Referent je sice nejasný, ale z kontextu můžeme aspoň částečně vymezit skupinu lidí (objektů), ke které odkazuje:

- ▷ *Jedním z hlavních témat řady rozhovorů dánského ministra zahraničí Nielse Helvega Petersena v Praze byla kromě bilaterálních vztahů i současná situace v Evropské unii. Dánská cesta k přijetí Maastrichtské smlouvy nebyla vůbec snadná: první celonárodní hlasování smlouvu odmítlo. Bylo zapotřebí řady měsíců* ***přesvědčování****, aby se v opakovaném* ***hlasování*** *Dánové nakonec těsnou většinou vyjádřili pro Maastricht.*

Z kontextu předpokládáme, že tím, kdo je přesvědčován, jsou Dánové. Koreferenční vztah je zde jasný, antecedent je explicitní: pro Adresáta verbálního substantiva *přesvědčování* volíme lemma *on*, od nějž vedeme k antecedentu *Dánové* koreferenční šipku. U Aktora je situace složitější: explicitní antecedent se v textu nevyskytuje, nicméně na základě kontextu je možné vyvodit, že jím jsou pravděpodobně politici. Nevolíme proto ani lemma *on* (antecedent není explicitně vyjádřen), ani lemma *Gen* (antecedent není všeobecný, je možné jej blíže specifikovat - vymezením skupiny „politici“), ale lemma *Unsp*. Obdobná je situace u Patienta: předpokládáme, že Dánové jsou přesvědčováni politiky o vstupu do Evropské unie, na místě je tedy lemma *Unsp*. Podobně postupujeme u substantiva *hlasování*: lemmatem Patienta i Aktora je *Unsp*.

Unsp **u sloves**

Stanovit kritéria pro rozlišení lemmat *Gen* a *Unsp* není snadné;[7] abychom valenčnímu doplnění s významem ACT u slovesa mohli přiřadit lemma *Unsp*, musíme zvážit tři zásadní hlediska:

1. typická povrchová realizace
    - sloveso se shoduje s nulovým podmětem v 3. os. pl. anim. (např. *Vypnuli* ***[Unsp.ACT]*** *proud*)

[3] [Daneš, 1971], [Panevová, 1973b], [Panevová, 1973a]

[4] Jde o linii *Gen – Unsp – on/ten*.

[5] Jde o doplněný uzel u sloves a jejich nominalizací.

[6] Vágnost tohoto elementu je reflektována už v jeho označení lemmatem *Unsp(ecified)*.

[7] V bohemistické tradici se také pro konstrukce s elidovaným Aktorem, jemuž přiřazujeme lemma *Unsp*, používá obdobný termín jako pro konstrukce s *Gen*.ACT, a sice věty se všeobecným podmětem.

- typická je také přítomnost adverbiálního určení místa, které lokálně vymezuje skupinu lidí, mezi nimiž předpokládáme daného referenta:
  - ▷ *Na poště.LOC zavírají **[Unsp.ACT]** v šest hodin odpoledne.*
  - ▷ *Tady.LOC dobře vaří **[Unsp.ACT]**.*

2. možnost vymezit referenta
3. vyloučení mluvčího ze skupiny možných Aktorů

Unsp jako lemma Aktora slovesa

Rozdíly mezi lemmatem *Gen* a lemmatem *Unsp* u ACT jsou shrnuty v tabulce 5.1.

Lemma uzlu	Vyloučení mluvčího	Typická povrchová realizace	Vymezení referenta
Gen	nevíme	reflexivní pasivum Př.: *Tato potvrzení se vydávají **[Gen.ACT] [Gen.ADDR]** na počkání.*	všichni Aktoři typičtí pro danou situaci
Unsp	ano	sloveso je ve tvaru 3. os. pl. anim. Př.: *Ukradli **[Unsp.ACT]** nám auto.*	skupina lidí není vymezená explicitně, ale můžeme vyvodit možného referenta z kontextu

Tabulka 5.1: Rozdíly mezi lemmatem *Gen* a lemmatem *Unsp* u ACT.

▷ *Zmizení tohoto 700 kg těžkého lékařského přístroje ... hygienikům ohlásili **[Unsp.ACT]** 30. června letošního roku. Podle informací LN však zářič ze skladu Škody Plzeň zmizel již koncem loňského roku.*

▷ *Co jste dělal mezitím? Začít tehdy samostatně režírovat na Barrandově bylo absolutně nemyslitelné. Ale přijali **[Unsp.ACT]** mě do scénaristického oddělení – nejdříve na rok, pak natrvalo.*

Jiná valenční doplnění sloves

Lemmatem *Unsp* může být označen kterýkoli aktant valenčního rámce slovesa; u těchto doplnění se však nemůžeme řídit formou, jediným společným kritériem je možnost vymezit referenta. U ***Unsp* jako Adresáta** je navíc typická (a téměř nutná) přítomnost adverbiálního určení místa:

▷ *Doma.LOC slíbil **[Unsp.ADDR]**, že přijde brzy, ale kamarádům řekl něco jiného.*

▷ *Celý den o tom jednal **[Unsp.ADDR]** ve vládě.LOC.*

***Unsp* jako Patiens** je nejhůře vymezitelný typ. Jistou pomůckou pro odlišení *Gen* a *Unsp* je tato specifikace:

1. jde-li o gnómické, nadčasové užití lexému → *Gen*
2. jde-li o kontextové, aktuální užití → *Unsp*

▷ *Jakmile měla trochu času, už gruntovala, vynášela **[Unsp.PAT]**, přestavovala **[Unsp.PAT]**. Když se točila u plotny.LOC, míchala **[Unsp.PAT]**, přisypávala **[Unsp.PAT]**, přilévala **[Unsp.PAT]**.*

Nabízí se také možnost použít lemma *Unsp* u těch případů tranzitivních sloves, kde se téměř lexikalizovalo intranzitivní užití,[8] např. *smeknout, zaparkovat, utrácet, zapálit si, zavěsit.* Objektová pozice je zde sice jednoznačně doplnitelná, množina antecedentů je tedy vymezena (*zaparkovat **auto***, *utrácet **peníze***, *zapálit si **cigaretu***), ovšem v textu obvykle nenajdeme žádný antecedent.

Unsp u substantiv

Ani u doplnění substantiv se nemůžeme řídit formou a jediným společným kritériem zůstává opět možnost vymezit referenta (srov. příklady se substantivy *přesvědčování* a *hlasování* na začátku kapitoly 5.2.2).

Nabízí se však možnost použít lemma *Unsp* pro elidované aktanty u substantiv rozvitých adjektivem, u kterého není jasné, zda jde o aktant, nebo o doplnění s funkcí restriktivního přívlastku (RSTR):

- ▷ *soudní.RSTR rozhodnutí **[Unsp.ACT]***
- ▷ *izraelsko-palestinská.RSTR jednání **[Unsp.ACT] [Unsp.ADDR]***
- ▷ *členská.RSTR evidence **[Unsp.PAT]***
- ▷ *Bylo by možné tvrdit, že pozornost věnovaná oběma tvůrcům je přehnaná, že jejich umění je uměním skandálu, warholovským uměním **[Unsp.ACT]** upozornit na sebe.*

5.3 Změna lemmat

Protože je textová koreference především kontextovou záležitostí, mají anotátoři koreference jiný přístup k anotaci lemmat než anotátoři velkého souboru. Vzhledem k tomu, že jsou anotátoři velkého souboru instruováni doplňovat povrchově nevyjádřené lexémy způsobem, který se na kontextové zapojení příliš neorientuje, mění anotátoři koreference po pečlivé úvaze lemmata tak, aby bylo možno koreferenci zachytit v případech, kdy to kontext vyžaduje.[9] Jedná se o přepis doplněných lemmat *Gen* na lemmata *on* (popřípadě *ten*) či *Unsp* podle potřeby, a to u analytického pasiva a doplněných aktantů valenčních rámců sloves a verbálních substantiv.

Případy, kdy je nutné se rozhodnout, zda jde o textovou koreferenci, anebo o vyjádření se všeobecným konatelem, hodnotí anotátor především podle kontextu; postupuje však podle těchto základních pravidel:

5.3.1 Analytické pasivum

Přepis lemmatu *Gen* na *on* je na místě pouze v případech, kdy je z kontextu jasné, kdo je konatelem slovesa pasivní konstrukce:

- ▷ ***Letectvo a zachránci** z britské pobřežní stráže včera uskutečnili dvě úspěšné operace a evakuovali cestující a posádky ze dvou lodí, které začaly hořet nedaleko britských břehů. Přes 100 lidí **bylo zachráněno [Gen.ACT]** z trajektu Sally Star.*
  (*Gen.ACT se přepíše na on.ACT, jehož antecedentem je letectvo a zachránci z předchozí věty)*

Časté jsou případy, kdy je vhodnější zvolit lemma *Unsp*:[10]

- ▷ *Ačkoli řada snímků českých oper na různých značkách - reedic i prvých vydání vůbec - byla už v LN recenzována **[Unsp.ACT]**, jistě nebude na škodu pohled zvnějšku . . . Na světovém trhu jsou v současnosti dvě nahrávky Dvou vdov, Krombholcova a Jílkova. Zatímco ostatní sólisté jsou na obou snímcích hodnoceni **[Unsp.ACT]** obdobně, Zahradníček se údajně nemůže měřit se svěžím zápalem Miroslava Švejdy a jeho elegantní tónovou linkou. . .*

[8] [Štícha, 1987]
[9] Anotování PDT probíhá na několika úrovních: kontextové zapojení je zohledňováno mimo anotaci velkého souboru.
[10] Srov. kapitolu 5.2.2.

5.3.2 Verbální substantiva

Aktanty valenčního rámce substantiv vstupují do koreferenčních vztahů zejména tehdy, je-li dějová složka verbálního substantiva vnímána jako kontextově aktuální. Jistou (ne však všeobecně platnou pomůckou) je to, že v těchto případech můžeme jednotlivé části valenčního rámce realizovat také lexikálně – například substitucí zájmena *on* či *ten*.

U verbálních substantiv bývá kontextově zapojen především Aktor a Patiens:

▷ *Děti mají ze školy stále spíše strach, než aby měly radost z **nabývání [Gen.ACT]** vědomostí.* (Gen.ACT se přepíše na *on*.ACT. Antecedent: *děti*.)

▷ *Radikálnější z Jižanů mluvili na tiskové konferenci dokonce o tom, že jejich země byly **pozváním [Gen.PAT]** do Osla podvedeny* (Gen.PAT se přepíše na *on*.PAT. Antecedent: *jejich*.)

Naopak v případech, kdy se jedná o děj často se opakující, ne právě aktuální, koreferenci nezaznačujeme a ponecháváme původní lemmata všeobecných aktantů:

▷ *Změna zákona o zdravotním **pojištění***

▷ *Tento tuk je vhodný na **pečení**.*

▷ *Nahoře se již linky rozvětvují, provoz řídne a **provětrávání** je lepší, takže trolej může skončit a pohon vozidla by převzal spalovací motor.*

▷ *Podobných vhodných míst by se našlo po naší republice hodně, zejména v městech s velkými převýšeními, která byla bývalým federálním ministerstvem dopravy vytipována k **zavedení** trolejbusové dopravy.*

5.3.3 Slovesné valenční rámce

Povrchově nevyjádřené aktanty valenčního rámce slovesa vstupují do koreferenčních vztahů poměrně často. Podle toho, o jaké kontextové zapojení se jedná, volíme lemma *on*, *Unsp* či *ten*. Například u slovesa *odmítat* v následujícím příkladu je možné doplnit na pozici Gen.PAT zájmeno *ten*:

▷ *„Některé ženy, převážně po čtyřicítce, si přejí doživotně zvýraznit rty, obočí nebo linky kolem očí. Jednoznačně odmítáme **[Gen.PAT]**“, dodává ještě Zdeněk.*

▷ *Kluk odpočívá dlouze, nehnutě na jedné noze. Když přijdu blíž, postaví se na obě a řekne **[Gen.ADDR]** mile: džambo, memsahib, how do you do?* (Gen.ADDR ⟶ on.ADDR ⟶ antecedent: *já*)

Pokud lze vyjádření chápat jako obecnou postulaci nebo je vhodnější je vnímat bez kontextového zapojení jako výsledek snahy mluvčího vyjádřit se obecně či neosobně, upřednostňujeme zachování tvarů lemmat jako všeobecných aktantů.

5.4 Nezaznačení textové koreference

Textovou koreferenci nezaznačujeme v následujících případech:

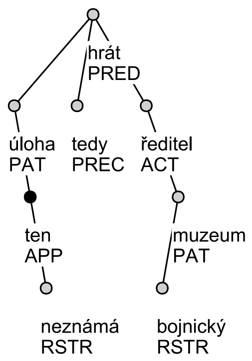
- **Frazémy, ustálená slovní spojení**
  ▷ *Tak je **tomu** i v těch případech, kdy dosavadní domovníci užívali byty na základě dohod s bytovými podniky nebo domovními správami, podle kterých jim byl přidělen byt po dobu výkonu domovnických prací.*

▷ *Mezitím do Pchanmundžomu, odkud byli v dubnu vypuzeni pozorovatelé České republiky, přijíždí i mnoho Korejců a hledí nepřítomně do dálky, na sever. Moc **toho** ovšem v tomto prostoru k vidění není.*

- **Intenzifikátory**
  - ▷ ***To** ale prší!*
  - ▷ *Ale **ono** je jedno, kdo dá gól, důležité je vyhrát; a jméno střelce, to je až na druhém místě.*

- **Obsahově vyprázdněné (nadužívané) výskyty zájmena v přímé řeči**
  Jde zejména o případy, kdy má zájmeno zdůrazňovací funkci či plní funkci slovní vaty, proto pro ně nelze najít antecedent.[11]
  - ▷ *„Nedokáži teď odhadnout dopad zákona na Úřad pro vyšetřování, ale myslím, že **to** rozhodně neztíží jeho práci nějakým markantním způsobem," řekl Ruml.*
  - ▷ ***To** máte těžké, mladému to beztak ošlápnul táta.*
  - ▷ *. . . jak si už dlouho představuje její cestu do ciziny, do Španělska nebo Řecka, kam ji **to** táhne.*

- **Technické uzly**
  Zájmeno je doplněno do stromu jen ze strukturních důvodů, nemá žádný antecedent. Jde o zájmeno *ten* na místě elidovaného substantiva, které řídí shodný přívlastek vyjádřený adjektivem (např. obr. 5.2).



Obrázek 5.2: *Úlohu **neznámé** bude tedy hrát ředitel bojnického muzea.* (lt20#15)

5.5 Udržování koreferenčních řetězců

U gramatické koreference se řídíme gramatickými pravidly; u textové koreference zachováváme koherenci textu a udržujeme koreferenční řetězec:

- je-li antecedent ve stromě rozložen do více na sobě závislých uzlů, odkazuje se k nejbližšímu předcházejícímu uzlu;
- existuje-li možnost výběru mezi antecedentem a postcedentem, upřednostňujeme antecedent (později by celé propojení jednotlivých částí koreferenčního řetězce zajistilo rozšířené pojetí koreference);

[11] Srov. např. Šmilauerův „zdánlivý podmět/předmět" ([Šmilauer, 1947]).

- je-li nutné volit mezi dvěma antecedenty, které rozpojují koreferenční řetězec, volíme ten, který je více vlevo (zohledňujeme řešení vhodné pro aktuální členění věty);
- existuje-li možnost výběru mezi dvěma různě lexikálně realizovanými antecedenty, u nichž není na základě stávajícího pojetí koreference možné realizovat znázornění vzájemné referenční totožnosti, odkazuje se k nejbližšímu z nich (i zde počítáme s postupným propojením jednotlivých částí koreferenčního řetězce v rozšířeném pojetí koreference).

Kapitola 6

Anotovaná data

6.1 Základní údaje

V následujícím oddíle uvedeme základní kvantitativní vlastnosti dat, která byla anotována před dokončením této zprávy:

- Počet anotovaných souborů: 269
- Počet vět: 14 036 (v průměru 52 vět na soubor)
- Počet tektogramatických uzlů: 168 836 (v průměru 12 uzlů na větu)
- Počet koreferujících uzlů: 10 738 (6,36 % všech uzlů, 0,77 koreferujícího uzlu na větu)
- Uzly s gramatickou koreferencí: 5160 (48,05 % všech koreferujících uzlů)
- Uzly s textovou koreferencí: 5578 (51,95 % všech koreferujících uzlů)

Tabulky 6.1, 6.2, 6.3 a 6.4 obsahují nejčastější lemmata a funktory koreferujících uzlů (v závislosti na typu koreference).

6.2 Mezianotátorská shoda

Mezi měřitelné vlastnosti anotovaných dat by měla patřit i jejich kvalita. Co se týká anotování koreference v češtině, s nejvyšší pravděpodobností neexistují v současnosti žadná jiná data, která bychom mohli považovat za „zlatý standard“ a použít je pro vyhodnocení kvality našich anotací. Jediným způsobem, jak jsme tedy mohli získat alespoň přibližný obrázek o míře spolehlivosti ruční anotace podle našeho schématu, bylo vytvořit nezávisle na sobě dvě anotace týchž vět a porovnat je navzájem.

S cílem zjistit mezianotátorskou shodu byl proveden následující experiment. Dvě anotátorky (jedna z nich začínající) dostaly nezávisle na sobě tři desítky souborů, deset souborů týdně. Celkové vyhodnocení je v tabulce 6.5.[1]

[1] Paralelní anotace srovnávala a rozdíly v nich vyhodnocovala Kateřina Černá.

trlemma	počet výskytů
který	1505 (29,17 %)
&Cor;	1166 (22,60 %)
se	1125 (21,80 %)
jenž	384 (7,44 %)
kdy	130 (2,52 %)
kde	129 (2,50 %)
což	100 (1,94 %)
sám	81 (1,57 %)
co	55 (1,07 %)
kdo	44 (0,85 %)

Tabulka 6.1: Nejčastější lemmata koreferujících uzlů – gramatická koreference (5160=100%).

funktor	počet výskytů
ACT	2590 (50,19 %)
PAT	691 (13,39 %)
APP	598 (11,59 %)
COMPL	362 (7,02 %)
LOC	247 (4,79 %)
BEN	168 (3,26 %)
TWHEN	141 (2,73 %)
ADDR	133 (2,58 %)
DIR3	39 (0,76 %)
DIR1	27 (0,52 %)

Tabulka 6.2: Nejčastější funktory koreferujících uzlů – gramatická koreference (5160=100%).

trlemma	počet výskytů
on	4443 (79,65 %)
ten	1070 (19,18 %)

Tabulka 6.3: Nejčastější lemmata koreferujících uzlů – textová koreference (5578=100%).

funktor	počet výskytů
ACT	3134 (56,19 %)
PAT	1135 (20,35 %)
APP	497 (8,91 %)
ADDR	206 (3,69 %)
EFF	111 (1,99 %)
LOC	95 (1,70 %)
DIR1	70 (1,25 %)
BEN	66 (1,18 %)

Tabulka 6.4: Nejčastější funktory koreferujících uzlů – textová koreference (5578=100%).

	1. týden (lm39–48)		2. týden (ln01–10)		3. týden (lm51–60)	
Souhrnné údaje:						
Počet vět celkem	508		538		521	
Počet uzlů celkem	5 796		6 810		6 769	
Počet koreferujících uzlů	375		316		463	
Počet chyb celkem	113		25		26	
Rozložení typů chyb:						
Textová koreference:	188	63	164	17	257	22
- koreferující *on*	150	44	124	8	214	15
- koreferující *ten*	36	17	40	9	43	7
- ostatní	2	2	0	0	0	0
Gramatická koreference:	187	50	152	7	206	4
- koreferující *se*	54	6	38	0	46	1
- koreferující *který*	44	7	35	0	47	1
- koreferující *&Cor;*	32	9	40	4	60	1
- ostatní	57	28	39	3	53	1

Tabulka 6.5: Vyhodnocení mezianotátorské shody na třech dávkách po deseti souborech.

Výsledky měření mezianotátorské shody interpretujeme takto:

- Vysoká mezianotátorská shoda potvrzuje, že anotační schéma je dostatečně popsáno (tedy že anotační instrukce poskytují oporu pro většinu rozhodnutí), aby mohla být zahájena anotace „ve velkém“.
- Shoda dosažená v druhém a třetím týdnu ukazuje, že učící křivka anotátora je velmi strmá.
- Rozdíl v mezianotátorské shodě v gramatické a textové koreferenci odpovídá předpokladu, že gramatická koreference je často jednoznačně určena samotnou stavbou věty, zatímco textová koreference může vyžadovat porozumění většímu kontextu, které je do jisté míry individuální.

Kapitola 7

Shrnutí a práce do budoucna

V předcházejících kapitolách byly zmíněny důvody pro anotování koreference v PDT, popsáno technické zázemí a pokyny pro anotaci a vyhodnoceny vybrané vlastnosti dosud anotovaných dat. V dalším snažení se budeme soustředit na následující cíle:

- upřesňování anotačních instrukcí (např. rozšíření gramatické koreference i na zachycování recipročních vztahů ve valenčních rámcích sloves a substantiv);
- dokončení anotace koreference ve všech větách PDT 2.0 (celkem přibližně 50 000 vět); koreference je v tuto chvíli anotována v jedné třetině dat;
- testování konzistence anotovaných dat;
- další vývoj softwarových nástrojů pro automatickou anotaci koreference.

Literatura

[Bruneseauxová and Romary, 1997] Bruneseauxová, F. and Romary, L. (1997). Codage des références et coréférences dans les dialogues homme-machine. In *Proceedings of the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, pages 15–17, Canada.

[Čmejrková, 1998] Čmejrková, S. (1998). *Issues of valency and meaning*, chapter Syntactic and discourse aspects of reflexivization in Czech: The case of the reflexive pronoun „svůj". Karolinum, Prague.

[Daneš, 1971] Daneš, F. (1971). Větné členy obligatorní, potenciální a fakultativní. *Miscellanea Linguistica*, pages 131–138.

[Fligelstone, 1990] Fligelstone, S. (1990). *A description of the conventions used in the Lancaster Anaphoric Treebank Scheme*. Departement of Linguistic and Modern English Language, Lancaster University.

[Gardentová et al., 2002] Gardentová, C., Manuélian, H., and Kow, E. (2002). Which bridges for bridging definite descriptions. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (Tag+6)*. Universitá di Venezia.

[Hajičová et al., 2000] Hajičová, E., Panevová, J., and Sgall, P. (2000). Coreference in annotating a large corpus. In *Proceedings of LREC 2000*, volume 1, Atény, Řecko.

[Hajičová et al., 2000] Hajičová, E., Panevová, J., and Sgall, P. (2000). A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank. Technical Report TR-2000-09, ÚFAL MFF UK, Prague, Czech Republic. in Czech.

[Halliday and Hasanová, 1976] Halliday, M. A. K. and Hasanová, R. (1976). *Cohesion in English*. Longman, London.

[Hirschman, 1997] Hirschman, L. (1997). MUC-7 coreference task definition. Version 3.0.

[Hoffmannová, 1997] Hoffmannová, J. (1997). *Stylistika a ...* Trizonia, Praha.

[Kořenský et al., 1987] Kořenský, J., Hoffmannová, J., Jaklová, A., and Müllerová, O. (1987). *Komplexní analýza komunikačního procesu a textu*. Pedagogická fakulta, České Budějovice.

[Koktová, 1992] Koktová, E. (1992). On new constraints on anaphora and control. *Theoretical Linguistics*, 18:102–178.

[Šmilauer, 1947] Šmilauer, V. (1947). *Novočeská skladba*. Praha.

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

[Orăsan, 2000] Orăsan, C. (2000). CLinkA a coreferential links annotator. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. ELRA.

[Panevová, 1973a] Panevová, J. (1973a). Všeobecný konatel a jeho vztah k mluvčímu. *Otázky slovanské syntaxe*, IV/1:101–106.

[Panevová, 1973b] Panevová, J. (1973b). Věty se všeobecným konatelem. *AUC – Studia Slavica Pragensia*, pages 133–144.

[Panevová, 1986] Panevová, J. (1986). K voprosu o refleksivnoj pronominalizacii v češskom jazyke. *Linguistische Arbeitsberichte*, 56(54).

[Panevová, 1991] Panevová, J. (1991). *Etudes de linguistique romane et slave*, chapter Koreference gramatická nebo textová? Kraków.

[Panevová, 1996] Panevová, J. (1996). *Prague Linguistic Circle Papers*, chapter More Remarks on Control, pages 101–120. 2. J. Benjamins Publ. House, Amsterdam — Philadelphia.

[Panevová, 1998] Panevová, J. (1998). *Tipologija, grammatika, semantika. K 65-letiju Viktora Samuilovit'a Chrakovskogo*, chapter Ellipsis and Zero Elements in the Structure of the Sentence, pages 67–76. Karolinum, Sankt-Peterburg.

[Panevová et al., 2002] Panevová, J., Řezníčková, V., and Urešová, Z. (2002). The theory of control applied to the Prague Dependency Treebank. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (Tag+6)*, pages 175–180. Universitá di Venezia.

[Petkevič, 1995] Petkevič, V. (1995). A new formal specification of underlying structures. *Theoretical Linguistics*, 21(1):1–61.

[Plátek et al., 1984] Plátek, M., Sgall, J., and Sgall, P. (1984). *Contributions to Functional Syntax, Semantics and Language Comprehension*, chapter A Dependency Base for a Linguistic Description, pages 63–97. Academia, Prague.

[Poesio and Vieira, 1998] Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite descriptions use. *Computational Linguistics*, 24(2):183–216.

[Poesio and Vieira, 2000] Poesio, M. and Vieira, R. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*. http://semanticsarchive.net/Archive/DZiNDg1M/poesio.cl2000.ps.

[Štícha, 1987] Štícha, F. (1987). Komunikativní a jazykové funkce lexikálního nevyjádření objektu děje ve větě. *Naše řeč*, 70:184–193.

[Tutinová et al., 2000] Tutinová, A., Trouilleux, F., Clouzotová, C., Gausier, E., Zaenenová, A., and Rayotová, S. (2000). Annotating a large corpus with anaphoric links. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, Lancaster, UK.

[Řezníčková, 2002] Řezníčková, V. (2002). PDT: Two Steps in Tectogrammatical Annotation with respect to some Issues of Deletion. *Prague Bulletin of Mathematical Linguistics*, 78:37–52.