Univerzita Karlova v Praze

Filozofická fakulta

Ústav českého jazyka a teorie komunikace

ROZŠÍŘENÁ TEXTOVÁ KOREFERENCE A ASOCIAČNÍ ANAFORA

(Koncepce anotace českých dat v Pražském závislostním korpusu)

EXTENDED NOMINAL COREFERENCE AND BRIDGING ANAPHORA

(An approach to annotation of Czech data in Prague Dependency Treebank)

Dizertační práce

Filologie – český jazyk

Anna Nedoluzhko

Školitel: prof. PhDr. Oldřich Uličný, DrSc.

2010

Prohlašuji, že jsem disertační práci vypracovala samostatně s využitím uvedených pramenů a literatury.

Anna Nedoluzhko

OBSAH

PODĚKOVÁNÍ

Poděkování za vedení při přípravě práce patří především mému školiteli prof. PhDr. Oldřichu Uličnému, DrSc., který mě neustále podporoval a povzbuzoval po celou dobu doktorského studia a při psaní této práce. Vděčím mu za mnohá důležitá poučení, detailní připomínky a v neposlední řadě za to, že mě naučil vážit si technické stránky práce a přesnosti v bibliografii.

Velký dík patří také prof. PhDr. Evě Hajičové, DrSc., která mě podporovala v záměru věnovat se zvolenému tématu, byla ochotná mi vždy poskytnout konzultace a dala mi v tomto směru řadu cenných rad.

Dále chci poděkovat Šárce Zikánové, PhD a Svatavě Škodové, PhD za ochotu a pečlivost, se kterou pročetly a okomentovaly mé přípravné verze. Vděčím jim a mým kolegům z Ústavu formální a aplikované lingvistiky na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze za mnohé důležité podněty, kterých se mi dostalo v diskusi.

V neposlední řadě bych ráda poděkovala anotátorům Radku Ocelákovi a Jiřímu Perglerovi za pečlivou a spolehlivou anotaci koreferenčních a anaforických vztahů na PDT, výborné nápady při řešení problematických případů, jejich dochvilnost a zodpovědnost při odevzdávání dat.

Za vytvoření a podporu anotačního nástroje a jeho přizpůsobení k potřebám anotace koreference děkuji Jiřímu Mírovskému, PhD a Petru Pajasovi, PhD.

Za finanční podporu tohoto výzkumu pak děkuji grantové agentuře GAČR, která jej podporovala v rámci grantu GAČR 405/09/0729 – Od struktury věty k textovým vztahům.

Na závěr patří největší dík mé rodině, která mě při napsání této práce maximálně podporovala. Mamince děkuji za to, že se po dobu psaní větší části práce věnovala mé malé dceři. Manželovi vděčím za pomoc s technickou stránkou práce, za dlouhé hodiny teoretických diskusí o možnostech automatického zpracování koreference a za kritický náhled na moje nápady. Dceři Alence děkuji za to, že mě naučila správně organizovat a produktivně užít pro práci sebemenší časovou skulinku.

ABSTRAKT

V této práci představujeme jeden z možných modelů zpracovaní rozšířené textové koreference a asociační anafory na velkém korpusu textů, který dále používáme pro anotaci daných vztahů na textech Pražského závislostního korpusu.

Na základě literatury z oblastí teorie reference, diskurzu a některých dalších poznatků teoretické lingvistiky na jedné straně a s použitím existujících anotačních metodik na straně druhé jsme vytvořili detailní klasifikaci textově koreferenčních vztahů a typů vztahů asociační anafory.

V rámci textové koreference rozlišujeme dva typy textově koreferenčních vztahů – koreferenční vztah mezi jmennými frázemi se specifickou referencí a koreferenční vztah mezi jmennými frázemi s nespecifickou, především generickou referencí.

Pro asociační anaforu jsme stanovili šest typů vztahů: vztah PART mezi částí a celkem, vztah SUBSET mezi množinou a podmnožinou/prvkem množiny, vztah FUNCT mezi entitou a unikátní funkcí na této entitě, vztah CONTRAST sémantického a kontextového protikladu, vztah ANAF anaforického odkazování mezi nekoreferenčními entitami a vztah REST pro jiné případy asociační anafory.

Jedním z úkolů výzkumu bylo vytvořit systém teoretických principů, které je nutno dodržovat při anotaci koreferenčních vztahů a asociační anafory. V rámci tohoto systému byl zaveden například princip důslednosti anotace, princip dodržování maximálního koreferenčního řetězce, princip kooperace se syntaktickou strukturou tektogramatické roviny, princip preference koreferenčního vztahu před asociační anaforou a další.

Vypracovanou klasifikaci jsme aplikovali na koreferenční a anaforické vztahy v Pražském závislostním korpusu (Prague Dependency Treebank, PDT). Byla provedena anotace těchto vztahů na polovině korpusu PDT (cca 25 tis. vět). Srovnání shody mezi anotátory při navazování vztahů a určování jejich typu ukázalo, že použitá klasifikace při daném rozsahu materiálu je spolehlivá zejména pro účely teoretického výzkumu; pro počítačové aplikační účely (strojový překlad, automatické učení atd.) je nutné rozšíření materiálové základny.

Klíčová slova

reference, koreference, anafora, textová koreference, asociační anafora, anotace, PDT

ABSTRACT

The dissertation presents one of the possible models of processing extended textual coreference and bridging anaphora in a large textual corpora, which we then use for annotation of certain relations in texts of the Prague Dependency Treebank (PDT).

Based, on the one hand, on the literature concerning the theory of reference, discource and some findings of theoretical linguistics, and, on the other hand, using the existing methodology of annotations, we created a detailed classification of textual coreferential relations and types of bridging anaphora.

Within textual coreference, we distinguish between two types of textual coreferential relations – coreferential relations between noun phrases with specific reference and coreferential relation between noun phrases with non-specific, primarily generic, reference.

We determined six types of relations for bridging anaphora: relation PART – between part and whole; relation SUBSET – between a set and a subset or element of a set; FUNCT – between an object and a unique function on that entity; CONTRAST – between semantic and contextual opposites; relation ANAF of anaphorical referencing between noncoreferencial objects; REST – for other examples of bridging anaphora.

One of the goals of the research is to create a system of theoretical principals that would be used for annotating coreferential relations and bridging anaphora. These principles include consistency of anaphora, the principle of maintaining maximum length of coreferential chains, principle of cooperation with the syntactic structure of the tectogrammatical tree, principle of preference of coreferencial relations over bridging anaphora and so on.

We applied the detailed classification to the coreferential and anaphorical relations in PDT. Comparison of the agreements between annotators during determination of relations and their types revealed that the classification was relatively reliable, although the inter-annotator agreement has not yet reached the significance that can qualify the annotated corpora of text for applied usage.

Keywords

reference, coreference, anaphora, textual coreference, bridging anaphora, annotation, PDT.

I. Úvodem

Pražský závislostní korpus 2.0 (Prague Dependency Treebank, dále PDT 2.0) je soubor velkého množství českých textů obohacených podrobnou a mezi sebou propojenou lingvistickou informací na morfologické, povrchově syntaktické a hloubkově syntaktické (tektogramatické) rovině.[1] Tektogramatická rovina, reprezentovaná jako orientovaný strom, který zachycuje hloubkovou strukturu věty, obsahuje rovněž informaci o aktuálním členění věty a některé druhy koreferenčních vztahů mezi uzly.

Existující anotace koreference v PDT 2.0 vychází z pojmu reference jazykových jednotek (referencí rozumíme vztah výrazů k předmětům nebo situacím reálného světa) a dělí se na gramatickou a textovou koreferenci. Gramatická koreference a některé případy pronominální textové koreference jsou již zpracovány na celém korpusu textů PDT. Anotace koreference je představena ve velkém manuálu Anotace na tektogramatické rovině Pražského závislostního korpusu (Mikulová a kol. 2005). Podrobný popis anotačního schématu, a to jak po stránce lingvistické, tak po stránce technické, je obsažen v (Kučová a kol. 2003) a v (Kučová – Hajičová 2004).

Tato dizertační práce vznikla v průběhu uskutečnění projektu kompletního zpracování rozšířené textové koreference a asociační anafory na tektogramatické rovině PDT. Během anotace se často vyskytují příklady, k jejichž řešení musíme využít teoretických poznatků z oblastí teorie reference, teorie anafory, kognitivní lingvistiky, teorie diskurzu aj. Tyto případy jsou natolik teoreticky zajímavé, že je třeba je zaznamenat a vysvětlit. Otvírají často celé oblasti zatím nevyřešených a velice zajímavých teoretických problémů. Na druhé straně, při analýze teoretické literatury vychází najevo, že některé doposud nevyřešené problémy mohou být vyřešeny pomocí analýzy většího korpusu textů, ve kterém jsou označeny koreferenční a anaforické vztahy. Tato práce je pokusem o spojení teoretického základu z oblasti především teorie reference s praktickým přístupem k anotaci koreference na velkém korpusovém materiálu, který vychází z principů aplikované lingvistické disciplíny – počítačové lingvistiky. Z toho důvodu práce obsahuje poměrně rozsáhlý přehled teoretických lingvistických zdrojů k teorii reference a anaforických vztahů, které nejsou všechny využity v praktické části práce. Řešení konkrétních anotačních a jazykových problémů často daleko přesahuje možnosti dané práce, proto na ně většinou pouze poukazujeme jednotlivými příklady, s tím, že je ponecháváme jako možnost pro budoucí rozpracování.

[1] Podrobně o PDT viz např. Mikulová a kol. 2005, Hajičová a kol. 2006.

Další zásadou této práce je snaha o propojení relativně podrobné lingvistické referenční analýzy a formálního aplikačního přístupu. V současné době existující anotace substantivní koreference a asociační anafory jsou buď založeny na velice obecných kritériích a mají lepší mezianotátorskou shodu (projekty MUC (Hirschman 1997) a ACE (Doddington a kol. 2004); MATE (Poesio a kol. 2000)) nebo naopak mají příliš specifická kriteria, která jsou špatně automaticky zpracovatelná (Passonneau 1996, rozšířená verze PoCoS (Charchos – Krasavina 2005) aj.). Ve zpracování koreference na textech PDT jsme se pokusili o kompromis – na jedné straně jsme vymezili základní typy, které se dají zpracovat přesnými metodami automaticky, na druhé straně jsme však typy koreferenčních a anaforických vztahů podrobili dostatečně detailní sémantické klasifikaci, která může přispět i pro řešení teoretických lingvistických úkolů.

Za koreferenční považujeme a jako koreferenční označujeme výrazy, které odkazují na tentýž objekt skutečnosti, pojem nebo situaci, přičemž platí, že koreferenční entity jsou vzájemně zaměnitelné bez věcné změny obsahu (stylistická změna je možná). Termín *koreference* implikuje pouze identitu referentů objektů, přesto občas pro zjednodušení výkladu o celém systému jevů odkazování používáme termínu koreference i pro případy nekoreferenčních vztahů, především pro mimotextové odkazování a asociační anaforu.

Následující dizertační práce má především praktické zaměření a je pojata jako návod anotace koreferenčních vztahů a asociační anafory na tektogramatické rovině PDT. Avšak klademe také důraz na teoretické vysvětlení konkrétních rozhodnutí, kterým dáváme přednost v problematických případech.

Celkově si tato práce klade za cíl na základě dokladů z korpusu PDT systematizovat a klasifikovat některé jevy z oblasti reference, koreference a anafory a připravit dostatečně reprezentativní ručně oanotovaný materiál vhodný k řešení následujících úkolů:

- Teoretický lingvistický výzkum:
  i. vlastnosti koreferenčních výrazů a anaforického odkazování;
  i. realizace anafory v textu, výběr konkrétního anaforického výrazu, elipsy, pronominalizace apod.;
  ii. heuristické výzkumy z oblasti aktivovanosti (salience);
  iii. kognitivní výzkumy, jako např. jak mluvčí využívají anaforické odkazy pro strukturaci informací v textu atd.

- Praktické aplikace:
  i. automatická generace výrazů v koreferenčních řetězcích;

ii. automatické rozpoznávání anafory (anaphora resolution);
iii. automatické rozpoznávání koreference (coreference resolution);
iv. automatické porozumění textu;
v. strojové učení;
vi. automatický překlad;
vii. dialogové systémy, automatické excerpce informace (information extraction);
viii. automatické odpovídání na otázky (question answering) a jiné NLP aplikace.

- Pro evaluaci výsledků při řešení uvedených výše úloh.

Práce je rozdělena do pěti oddílů. Oddíl II představuje teoretický kontext výzkumu anafory a koreference v oblastech kognitivní sémantiky a teorie diskurzu (II.1.), teorie reference (II.2.) a počítačové lingvistiky (II.3.). V kapitole III.1. jsou formulovány principy a preference anotace koreference. Kapitola III.2. popisuje formální charakteristiky koreferovaných uzlů. V oddíle III na základě použité literatury vytvořeno vlastní schéma anotace rozšířené koreference (III.4.) a asociační anafory (III.5.) na tektogramatické rovině v PDT.[2] Zvlášť se rozebírají problematické případy, kde se tyto dva jevy prolínají (III.6.) a speciální typy reference – mimotextové odkazování (III.7.1) a odkazování na větší úsek textu (III.7.2.). V oddíle IV. se pojednává a aplikaci probíhající anotace a jsou uvedeny první statistické a evaluační výsledky.

[2] Důsledné rozdělení koreference a asociační anafory je podmíněno tím, že tyto jevy jsou velmi odlišné a různě použitelné. Asociační anafora zatím nemůže být provedena automaticky. Vyčleňování asociační anafory jako anotovaného vztahu na velkém textovém korpusu není zcela zřejmé řešení (III.5., III.5.1.) – je to jev výrazně méně samozřejmý a spolehlivý. Je pravděpodobné, že v blízké budoucnosti nenajde uplatnění.

II. Lingvistický kontext výzkumu anafory a koreference

Anaforické odkazování a koreference jsou objektem výzkumu několika lingvistických disciplín. Zevrubně je můžeme rozdělit do čtyř skupin. První skupina je diskurzivní. Sem patří výzkumy z oblasti kognitivní lingvistiky (Schwarz-Frieselová 2007), textové lingvistiky, aktuálního členění (Averintsevová-Klischová – Consten 2007) a teorie diskurzu. Druhá skupina je teoreticky lingvistická. Do té spadají práce z oblasti syntaxe, sémantiky a teorie reference. Třetí oblast se týká počítačové lingvistiky. Do čtvrté skupiny patří neurolingvistický výzkum recepce anaforických vztahů.

Pro úplnost zpracování koreference a asociační anafory na velkém textovém korpusu je nutné nahlížet na dané téma z více hledisek. Na jedné straně se chceme opřít o lingvisticky filozofickou a logickou teoretickou tradici teorie reference a teoretické zpracování anafory (II.1. a II.2.). Na druhé straně náš přístup má být prakticky realizovatelný na velkém korpusovém materiálu, mít řešení pro konkrétní jazykové příklady, nejednoznačné případy, a zároveň nebýt přehuštěný typy a příznaky, abychom mohli dosáhnout rozumné mezianotátorské shody a aby anotace mohla být dokončena ve viditelném časovém horizontu a posloužit jako tréninkový materiál pro strojové učení a jiné možné aplikace. Z těchto důvodů potřebujeme využít existujících zkušeností z anotací anaforických a koreferenčních vztahů v projektech z oblasti počítačové lingvistiky aplikovaných na větší textové korpusy. O takových projektech pojednává kapitola II.3.

II.1. Kognitivní a diskurzivní přístupy k analýze anafory

Anaforické odkazování je jeden ze základních prostředků koherence textu. Tradičně se v textové lingvistice anaforickým vztahem rozumí pouhý odkaz na již uvedený v předcházejícím textu antecedent, se kterým je anafor rovněž koreferenční, přičemž základní funkce tohoto vztahu je dodržování tematické posloupnosti v textu. Srov. (1):

(1) *Helena poprosila maminku, aby na ni počkala.* *(VL)*[3]

Avšak ve skutečných textech je pojem anafory mnohem širší. Určitý a koherenční anaforický výraz může přinášet do textu novou nebo doplňující informaci, a v tom případě identifikace (budiž koreferenčního) antecedentu ve velké míře záleží na interpretačních schopnostech adresáta (viz o tom v Schwarz-Frieselová 2007, Schwarz 2000). Srov. např. (2):

[3] Zkratku *VL* používáme v textu pro příklady konstruované na základě vlastní řečové zkušenosti.

(2) *Nehodu zavinila řidička tramvaje. Třicetiletá Helena Beretová více než dvakrát převýšila povolenou rychlost. (VL)*

Podle Schwarz-Frieselové (2007) pro koreferenční interpretaci jmenných frází v kontextech typu (2) adresát vytváří jistý model daného textu (*text-world model*), který se v průběhu textu mění a doplňuje.[4] Kromě toho, při interpretaci textu tento model spolupracuje s modelem světa a systémem obecných znalostí adresáta.

Ani koreference není nutnou podmínkou anaforického vztahu. V případě, že postcedent a antecedent anaforického vztahu nejsou koreferenční, jde o asociační anaforu (*associative anaphors* u Hawkinse (1978), Vatera (1984) a Heima (1991)), nepřímou anaforu (*indirect anaphora* v Schwarz-Frieselové (2007)) nebo tzv. bridging anaforu (např. v Clarkovi 1977). V naší práci budeme používat termín *asociační anafora*. Srov. např. vztah mezi *strop* a *místnost* v (3) a *učitel, třída* a *děti* v (4):

(3) *Helena vstoupila do místnosti. Ze stropu kapala voda.*

(4) *Učitel vešel do třídy. Děti se okamžitě přestali bavit.*

Věty v (3) bezpochyby představují souvislý kontext, i když *strop* nemá v předchozí větě přímý koreferenční antecedent. Správná interpretace nepřímé anafory často vyžaduje zapojení kognitivních procesů, které zahrnují aktivaci veškerých sémantických sítí/struktur. Například pro správné pochopení (3) je nutné zapojit informaci o tom, že strop je částí místnosti, pro správnou interpretaci (4) je nezbytná znalost, že učitelé učí žáky ve třídě a tito žáci bývají děti.

Schwarz-Frieselová (2007) definuje nepřímé anafory jako „definite Nps, which have no explicit antecedent in text, but which are linked to some previously mentioned element by a cognitive process".[5] Autorka vyčleňuje následující základní charakteristiky asociační anafory:

- Anaforický člen nemá explicitní antecedent, avšak v kontextu mu předchází entita, se kterou nějakým způsobem souvisí, nebo ze které je vyvoditelný.

[4] Podobné koncepty jako *text-world model* u Schwarz-Frieselové najdeme také v mnoha jiných pracích, přičemž nejenom z oblasti kognitivistiky (srov. *денотативное пространство* (*denotační prostor*), *релевантное денотативное пространство* (*relevantní denotační prostor*) v teorii reference (Šmelev 1996), *průnik množin aktualizovaných znalostí* v teorii aktuálního členění Yokoyamové (2005), *aktivace, reaktivace a deaktivace referentů* v Givonovi (1992), *continuing, retaining* a *shifting* v „centering theory", např. v Maesovi (1997) aj.). Vysvětlení identifikace koherenčních elementů textu pomocí podobných pojmů jsou velice rozumná a intuitivní, používají se široce a liší se podle svého konkrétního zaměření.

[5] Schwarz-Frieselová a kol. (2007, s.IX).

- Mezi členy páru není vztah koreference, avšak jsou spojeny sémantickým nebo konceptuálním vztahem jiného druhu.
- Fungují jistá omezení na pronominalizaci anafora a na jeho použití s ukazovacím zájmenem.
- Identifikace anaforického vztahu v rámcích daného textu vyžaduje kognitivní proces včetně aktivace znalostních struktur.
- Na rozdíl od přímé anafory asociační anafora vždy přináší do modelu daného textu (*text-world model*) novou informaci. Přitom aktivační proces probíhá tak, že zároveň s tím, že se reaktivuje antecedent asociační anafory, v operativní paměti se otvírá nová „skříňka" pro referent asociačního anafora.

Asociační anafora není v žádném případě jev marginální a v souvislém textu se vyskytuje zcela běžně. Poukazují na ni různé prvky vyjadřující kontextovou zapojenost, topikalizaci nebo určenost v textu. Problém je však ve vymezení asociační anafory z řady případů přímé anafory, a na druhé straně její odlišení od případů definitivních deskripcí jiného druhu. Hranice na obou stranách má graduální charakter a má být vždy stanovena v souladu s požadavky konkrétního řešeného úkolu.

Interpretace anaforických výrazů ve velké míře závisí na kontextu – pro určení referentu anaforického výrazu adresát vyhledává informaci mimo anaforického výrazu samotného, přičemž znalosti, které pro tuto interpretaci potřebuje, mohou být různého druhu – znalost blízkého nebo vzdálenějšího kontextu, sémantických souvislostí a konceptů, nebo obecná znalost světa. Je pozoruhodné, že až na případy gramatické koreference (viz III.3.), kde určování antecedentu záleží na gramatických pravidlech jazyka, gramatická struktura textu je výrazně méně důležitá pro interpretaci anaforických souvislostí, než uvedené znalosti.

Dá se mluvit o široké škále typů asociační anafory. V každém případě pro interpretaci vztahu mezi antecedentem a nepřímým anaforem je zapotřebí použít jistý propojovací mechanismus. Interpretace asociační anafory může být založena na sémantice asociovaných členů nebo jejich vztah může být konceptuální povahy, přičemž v tom případě interpretace vztahu vyžaduje zapojení obecné znalosti světa, také různého druhu. Na základě těchto rozdílů Schwarz-Frieselová (2007, s. 8n.) vyčleňuje řadu nejčastěji se vyskytujících prototypických případů asociační anafory:

- ČÁST – CELEK (*auto – volant, člověk – noha*). Tento vztah se vyvozuje ze sémantiky zúčastněných substantiv. Znalosti, které jsou nutné pro interpretaci daného

anaforického výrazu, jsou uloženy v „mentálním slovníku" člověka (tzv. *mental lexicon*).

- SLOVESO – REALIZACE INTENČNÍ POZICE NEBO CIRKUMSTANTA[6] (*řídit – řidič, auto, zamknout – klíč* apod.). Podobně jako předchozí vztah, tento vztah se také vyvozuje ze sémantiky zúčastněných členů.
- Tzv. „frame-evoked entities" (*restaurace – jídlo, číšník*). Interpretace daného vztahu vyžaduje znalost konceptu a je založena na aktivaci znalosti „scénáře", např. *restaurace*. Pro správnou interpretaci je nutné umístit referent nepřímého anafora do mentálního prostoru scénáře.
- Nepřímé anafory koncepčního typu vyžadující dodatečnou indukci. Pro interpretaci vztahu tohoto typu není vždy zapotřebí aktivace znalostního konceptu nebo scénáře, ale adresát musí prodělat netriviální myšlenkový krok. Srov. interpretaci určité jmenné fráze *the rake* v (5):[7]

(5) *One night a man rushes into the police station and tells the policeman that he has just been knocked down in his garden. One policeman is asked to go and look for traces at the place of the assault. After a short time he returns with a huge swelling at his head and says "I solved the case". "Bravo", says his boss, "and how did you do that?" "I stepped on the rake, too."*

Jiná forma anaforického odkazu je tzv. komplexní anafora (*complex anaphora*). Podle Constena (Consten – Knees – Schwarz-Frieselová 2007), komplexní anafory jsou jmenné fráze, které odkazují na propozičně strukturované objekty (propozice, událost, stav, tvrzení).[8] V případě komplexní anafory mluvčí z části diskurzu vytváří metatextový abstraktní objekt (complexation process), na který se potom odkazuje. Zároveň s aktem reference úseku diskurzu může proběhnout akt evaluace antecedenční propozice. Srov. $(6)_{c,d}$:[9]

[6] "... Indirect anaphora based on [...] thematic roles of the verb in the preceeding sentence" (Schwarz-Frieselová 2007, s. 9).

[7] Příklad viz Schwarz-Frieselová 2007, s. 10.

[8] Viz definici v Consten – Knees – Schwarz-Frieselová (2007, s. 82): „Complex anaphors are nominal expressions referring to propositionally structured referents (such as propositions, states, facts and events) while introducing them as unified entities into a discourse. Additionally, they can classify or evaluate the referent."

[9] Příklad (6) je paralelní českou větou k příkladu uváděnému v Consten – Knees – Schwarz-Frieselová (2007, s. 82).

(6) *Mladí řidiči často jezdí velice rychle. To$_a$/Tato$_b$ skutečnost/Tento stereotyp$_c$/Tato neslušnost$_d$...*

Studie Marx – Bornkesselová-Schlesewsky – Schlesewsky (2007) prokázala, že proces identifikace a rezoluce komplexní anafory vyžaduje větší kognitivní úsilí, než identifikace a rezoluce anaforických jmenných frází. Odkazují totiž k abstraktním objektům (kterými jsou propozice, skutečnosti, události apod. realizované celou větou, nebo ještě větším úsekem textu), mohou být zařazeny do různých ontologických kategorií a motivují integraci nového abstraktního referenta do mentální reprezentace diskurzu.

Prototypická komplexní anafora je přímá, tj. pro identifikaci propozičního antecedentu je dostatečné vyhledat správný úsek předchozího textu. To však není obligatorní (k asociační komplexní anafoře viz Consten – Knees – Schwarz-Frieselová (2007) a příklady v sekci III.8.).

II.2. Anaforické vztahy v kontextu teorie reference

Výzkum anaforických mechanizmů jazyka (pronominalizace, elipsa, proces pojmenování objektů v různých pozicích v textu z hlediska anafory), textové koheze, jiných aspektů textové syntaxe atd. nebude dostatečně úplný a přesný, pokud nebude využívat poznatků z teorie reference. Teorie reference pramení v logice (srov. např. Frege 1892, Searle 1969 – klasická teorie reference, chápání reference jako identifikace výrazů s obecným významem, reference je dána především smyslem výrazu; Kripke (1980), Donnellan (1966, 1979) – neokauzální teorie reference vycházející hlavně z vlastních jmen a aplikující jejich způsob reference na jiné typy výrazů) a jejím předmětem je vztah jazykových výrazů k označované skutečnosti. Pro účely této práce jsme využili teoretické prameny především české a ruské tradice výzkumu reference. Také byly zohledněny některé německojazyčné slavistické práce (Berger 1993; Mendozová 2004, Weiss 1978, 1983 aj.).

II.2.1. Ruská tradice teorie reference

Ruská tradice teorie reference je představena především v pracích E. Padučevové (1985, 1979), E. Aruťunovové (1976) a A. Šmeleva (1996). Ve studiích Padučevové (1979, 1985) a Aruťunovové (1976) je na základě logicky sémantických kritérií vypracován systém tzv. denotačních statusů (typů reference) substantiv (hlavně) s předmětným významem v různých pozicích.

Reference je vlastnost, kterou jazykové výrazy nabývají v konkrétním textu (výpovědi, diskurzu). Souvislost mezi jazykovým významem a referencí je různá u různých typů výrazů schopných reference. Na základě tohoto kritéria **Padučevová** (1985, s. 81) dělí jména na čtyři skupiny:

- vlastní jména, u kterých reference je založena na pragmatických znalostech mluvčího o světě,
- deiktické výrazy, které nabývají (vždy stejného) významu v konkrétním mluvním aktu,
- deskripce (kombinace obecného jména a deiktických výrazů),
- obecná jména, která nemají vlastní referenci a nabývají ji pouze v kombinaci s deiktickými výrazy.

Jmenná fráze se podle Padučevové skládá z obecného jména (*общее имя*) a aktualizátoru (*актуализатор*). Obecné jméno je slovníková jednotka, která v případě předmětného jména má tzv. extensionál (množinu objektů, které může označovat daná NP) a není zapojena časově a lokálně. Aktualizátory jsou jazykové jednotky (slova nebo komponenty věty), které činí z obecného jména aktualizovanou NP, zapojenou do kontextu a obsahující časoprostorové charakteristiky. Jako aktualizátory mohou sloužit např. různé typy zájmen (*ten, takový, každý, nějaký* apod.). Aktualizátor může mít nulovou hodnotu, tj. nemusí být ve větě explicitně vyjádřen (*Lékař přišel až večer.*[10]), na druhé straně se však aktualizovaná NP může skládat pouze z aktualizátoru bez obecného jména, jak je to v případě substantivních použití zájmen (např. *já,* substantivní *to* apod.), nebo se aktualizátor napojí na již aktualizovanou NP (srov. např. *některý* v *některý z těch studentů*). Referenční hodnota jmenné fráze se určuje především na základě významu aktualizátoru dané NP, přičemž u různých druhů NP se jejich referenční hodnota určuje s různou mírou jednoznačnosti (např. *anglický král* – vždy závisí na kontextu, vlastní jména – jediná interpretace), srov. Padučevová (1985, s. 84).

Klasifikace denotačních statusů podle Padučevové vyčleňuje především tzv. termové (substantivní, věcné, předmětné) použití NP od použití predikativního. Při predikativním použití NP se neprovádí reference na objekt, avšak jinému objektu se připisuje určitá vlastnost. Stává se to v případech, kdy jmenná fráze je součástí přísudku nebo se nachází v apozici. Srov. např. NP *lékař* v (1) a NP *krasavice* v (2):

(1) rus. *Иван врач.*
č. *Ivan je lékař.*

[10] Pokud není uvedeno jinak, příklady (1)–(21) pochází z Padučevové (1985), překlad do češtiny – *AN+ŠZ.*

(2) rus. *У него была дочь красавица.*
č. dosl. *Měl dceru krasavici.*

Zvlášť se vyčleňuje skupina tzv. autonymních použití, kde NP má pokleslý referent, jako např. *Nataša* v (3):

(3) rus. *Муж просто звал ее Наташей.*
č. *Manžel ji říkal prostě Nataša.*

Za predikativní se považují také NP typu *реки, которые зимой замерзают* (*řeky, které jsou v zimě pod ledem*) v (4), které by logicky mohly mít spíše existenciální status (viz dále).

(4) rus. *Есть реки, которые зимой замерзают.*
č. *Jsou řeky, které jsou v zimě pod ledem.*

Termové NP se dále rozdělují na jména se specifickou referencí (tzv. konkrétně-referenční, singulativní) a nereferenční jména. Jména se specifickou referencí individualizují objekt, např. v (5) referent NP *okno* je již vybrán z množiny všech významů lexémů *okno* a představuje určité okno v určitém místě.

(5) rus. *Окно было маленькое и узкое.*
č. *Okno bylo malé a úzké.*

Uvnitř třídy jmenných skupin se singulativní referencí se NP dále klasifikují na základě rysu „± určenost" (určenost objektu zároveň pro mluvčího a adresáta) a „± slabá určenost" (určenost jenom z hlediska mluvčího).

Silná určenost NP souvisí s presumpcí existence a jedinečností objektu ve společném kontextu mluvčího a adresáta, tj. mluvčí vždy předpokládá, že adresát bude daný výraz určitým způsobem interpretovat. Jedinečnost přitom může vyplývat z významu obecného jména (srov. (6)), je součástí aktualizátoru, pokud se na objekt přímo ukazuje, jako např. na knihu v (7) nebo (8), kde je prezumce existence jediné knihy, kterou mluvčí dostal od adresáta, nebo jsou jiné příčiny pro jednoznačnou identifikaci objektu, jako např. existence objektu ve společném *obzoru* mluvčího a adresáta v (9) apod.

(6) rus. *Лучшая из моих картин находится в Лувре.*

č. *Můj nejlepší obraz je vystaven v Louvru.*

(7) rus. *Я прочел эту книгу.*
č. *Přečetl jsem tuto knihu.*

(8) rus. *Ту книгу, которую ты мне дал, я уже прочел.*
č. *Knihu, kterou jsi mi dal, už jsem přečetl.*

(9) rus. *Если ты подойдешь ближе к компьютеру, я покажу тебе его фотографию.* (VL)
č. *Pokud se přiblížíš k počitači, ukážu ti jeho fotku.*

Určenost může být mimotextová a textová, přičemž pro textovou určenost není podmínkou bezprostřední použití dané NP v předchozím kontextu, ale může vzniknout i situací, kterou text generuje, srov. určenost NP *дорога* (*silnice)* v kontextu (10):

(10) rus. *Он возвращался домой поздно. Дорога была плохо освещена.*
č. *Vrácel se domů pozdě. Silnice byla špatně osvícena.*

Neurčité NP jsou v klasifikaci Padučevové dále rozděleny podle rysu „+/– slabá určenost". Slabou určenost ukazuje Padučevová na příkladech typu (11), kde adresát pravděpodobně nebude vědět, o kterou cizinku jde, zatímco pro mluvčího je objekt známý.

(11) rus. *Он хочет жениться на одной иностранке.*
č. *Chce se oženit s jednou cizinkou.*

Jmenné fráze s rysem „– slabá určenost" jsou neurčité také pro mluvčího, srov. např. (12):

(12) rus. *Иван читает какой-то учебник.*
č. *Ivan čte nějakou učebnici.*

Rys „+ slabá určenost" bývá často neutralizován, jestliže z kontextu nepoznáme, je-li daná jmenná fráze pro mluvčího určitá, srov. *милиционер* (*policista*) v (13):

(13) rus. *Иван подрался с милиционером.*

č. *Ivan se popral s policajtem.*

Nereferenční NP referují k nevybranému mimojazykovému objektu. Padučevová vyčleňuje nereferenční NP atributivní, univerzální, existenciální a generické. Atributivní NP lze ilustrovat např. jmennou frází *Убийца Смита* (*vrah Smitha*) v (14); při této interpretaci mluvčí předpokládá existence jediného vraha, kterého však nekonkretizuje (přibližně 'ten, kdo zavraždil Smitha, je blázen').

(14) rus. *Убийца Смита сумасшедший.*
č. dosl. *Vrah Smitha je blázen.*

Srov. také některé NP ve větách obecné povahy jako např. (15)–(16):

(15) rus. *Самый сильный человек в мире не в состоянии поднять больше 200 кг.*
č. *Ani ten nejsilnější člověk na světě nedokáže zvednout víc, než 200 kg.*

(16) rus. *Тот, кто победит в этой борьбе, не избежит нечестных приемов.*
č. *Ten, kdo vyhraje ten boj, se nevyhne nepoctivým postupům.*

Součástí významu univerzální NP je velký kvantifikátor, tj. tyto výrazy označují všechny objekty abstraktní třídy, která je extenzí daného jména, např. (17):

(17) *ru. Все дети любят мороженое.*
cz. Všechny děti mají rády zmrzlinu.

Existenciální NP se používají v situaci, když se mluví o objektu, který patří do množiny objektů podobného typu a přitom není individualizován, tj. není vybrán z dané třídy. Padučevová vyčleňuje tři typy existenciální reference: distributivní, nekonkrétní a obecně-existenciální. Distributivní NP (např. (18)) referují k účastníkům různých stejnorodých situací.

(18) rus. *Иногда кто-нибудь из нас его навещает.*
č. *Občas ho někdo z nás navštíví.*

Nekonkrétní jsou takové NP, které vystupují v kontextech bez afirmativity. Padučevová uvádí seznamy typů kontextů, které vytvářejí podklad pro takové použití. Jsou to např. modální

slova typu *může, chce, musí* aj.; rozkazovací způsob, budoucí čas, otázka, negace (včetně negace uvnitř lexému: *odmítat, nezbytně, zakázat*), disjunkce, podmínka, cíl, nejistota, předpoklad; některé propoziční predikáty: *chtít, myslet* aj. a performativní slovesa (*prosím, slibuji*). Srov. např. (19) v případě, že se příslušné osoby ještě neseznámily, a (20):

(19) rus. *Джон хочет жениться на какой-нибудь иностранке.*
č. *John se chce oženit s nějakou cizinkou.*

(20) rus. *Он ищет новую секретаршу.*
č. *Hledá novou sekretářku.*

Obecně-existenciální NP se vyskytují v textu, jestliže se mluví o objektech s určitými společnými vlastnostmi a referují např. k neurčitému počtu objektů určité třídy, aniž by tyto objekty byly individualizovány. Srov. (21):

(21) rus. *Некоторые вещи портятся при перевозке.*
č. *Některé věcí se mohou poškodit při stěhování.*

Do nereferenčních NP Padučevová (1985) zařazuje rovněž generické NP. Schematicky systém denotačních typů podle Padučevové shrnujeme v následující tabulce č. 1:

substantivní	referenční	+určitost	*Přečetl jsem tuto knihu.*	
		- určitost	+ slabá určitost	*Chce se oženit s jednou cizinkou.*
			- slabá určitost	*Ivan čte nějakou učebnici.*
	nereferenční	atributivní	*Vrah Smitha je blázen.*	
		univerzální	*Všem dětem chutná zmrzlina.*	
		existenciální	distributivní	*Občas ho někdo z nás navštíví.*
			nekonkrétní	*Hledá novou sekretařku.*
			obecně-existenciální	*Některé věcí se mohou poškodit při stěhování.*
		generické	*Štír vypadá jako cvrček.*	
predikativní	*Ivan je lékař.*			

Tabulka č. 1: Systém referenčních typů podle Padučevové

Klasifikace denotačních typů podle Padučevové je velice důsledná a pokrývá velkou část případů vyskytujících se v textu. Problém je však v tom, že daná klasifikace je orientována pouze na předmětná jména a neřeší referenci jmenných frází s méně konkrétním významem, které se v korpusu objevují minimálně stejně často jako předmětná jména. V článku (Padučevová 1986) se autorka sice věnuje problematice reference jmenných frází s nepředmětným významem, ale soustředí se hlavně na propoziční komponenty výpovědi a řeší je zcela jinak, než referenci předmětných jmen. Tvrdí totiž, že referentem propozičního komponentu významu je situace, kterou daná propozice popisuje. Takové chápání však vyvolává zcela odlišný přístup k referenci propozičních komponentů, což při zpracování velkého korpusu textů představuje neřešitelný problém – nemohli bychom sjednotit analýzu předmětných a nepředmětných jmen. Vzhledem k tomu, že hranice mezi předmětnými jmény, abstraktními jmény a jmény pojmenovávajícími situaci je spíše graduální povahy (viz III.4.2.2.), aplikace tohoto rozlišení při zpracování velkého korpusu textů by vyžádala velice složitý systém pravidel a konvenci, kterou si v dané etapě nemůžeme dovolit. Další komplikace, na kterou se narážíme při aplikaci klasifikace Padučevové na velký textový korpus, je, že v ukázkách typů reference Padučevové nejsou anaforické páry, ale jednotlivé výpovědi. Proto např. není jasné, do kterého referenčního typu se podle Padučevové zařadí anaforická jmenná fráze, kde antecedent má nespecifickou referenci a anafor je použit s explicitním determinátorem (např. s ukazovacím zájmenem).[11] Z ústní diskuze s prof. Padučevovou a z její klasifikace typu určenosti v článku (Padučevová 1985) vyplývá, že se autorka drží názoru, že pokud daná jmenná fráze má určitý identifikátor, nemůže být již pojata jako nespecifická (určenost je podle Padučevové příznak, který je aktuální pouze pro jmenné fráze se specifickou referencí). Při anaforickém odkazu s určitým determinátorem se totiž vytváří fiktivní model textu, ve kterém se daná jmenná fráze už chápe jako specifická. Připouští se však anaforický odkaz s determinátorem na generickou jmennou frázi, aniž by přestala být generickou,[12] což tomu tvrzení částečně odporuje.

V práci **A. D. Šmeleva** (Šmelev 1996) jsou referenční mechanismy jazyka popsány z poněkud odlišného hlediska. Ve svém popisu A. Šmelev vychází ze dvou základních pojmů – denotační prostor (*денотативное пространство*) a relevantní denotační prostor (*релевантное денотативное пространство*). Denotační prostor je jakýkoli úsek mimojazykové skutečnosti. Pro kterýkoli jazykový výraz, který se používá v promluvě, je

[11] Viz příklady v III.4.2.1.1. h).

[12] Viz Paducevová 1985, s. 112.

relevantní ten denotační prostor, ve kterém je určen referent daného jazykového výrazu.[13] Například relevantním denotačním prostorem pro NP *ректор* (*rektor*) v (22) je v případě přirozené interpretace denotační prostor univerzity, ve které se nachází účastníci komunikace.

(22) rus. *Я расскажу об этом ректору.*[14]
č. *Řeknu to rektorovi.*

Rozbor referenčních mechanizmů v práci Šmeleva je proveden v rámcích tzv. neokauzální teorie reference, která se zakládá na postupném obohacování tzv. *pomyslné složky* (*мысленного досье*) adresáta na daný referent. Autor to ukazuje na příkladu vlastních jmen, kde na začátku konverzace adresát nemusí o referentu, označeném daným jménem, vědět nic, potom však postupně dostává další a další informace, čímž se jeho pomyslná složka doplňuje. Rozvíjí se tedy kauzální řetězec od introduktivní promluvy (typu *To je Karel*) hlouběji do dalších a dalších informací o adresátovi.[15]

Mimojazykové objekty se dělí podle Šmeleva na třídy (*классы*) a prvky (*индивиды*). Třída je otevřená nepočetná množina objektů, prvky jsou jednotlivé objekty nebo uzavřené množiny objektů. Podle toho se typy reference dělí na generickou (*генерализованную*) a individuální (*индивидную*). Zvlášť se vyčleňuje hraniční oblast abstraktně-individuální reference. Referent takové NP je sice individuální, ale je použit ve výpovědi, která nemá časoprostorové charakteristiky, např. NP *pes* v (23)[16] nebo v pozici objektu syntaktických konstrukcí typu *пойти в магазин* (*jít do obchodu*), *лечь в больницу* (*jít do nemocnice*):

(23) rus. *Собака любит Ивана.*
č. *Pes má Ivana rád.*

Uvnitř generické reference se dále vyčleňují obecně generické (*общеродовые*) a obecněexistenciální (*общеэкзистенциальные*) jmenné fráze.

Obecně generické jmenné fráze odrazují k celé třídě objektů. Srov. např. referenci *pes* a *děti* v (24)–(25):

[13] Viz Šmelev 1996, s. 23.

[14] Příklady (22)–(29) jsou převzaty z (Šmelev 1996).

[15] Ibid. s. 33.

[16] Ibid. 43n.

(24) rus. *Собака – друг человека.*
č. *Pes je přítelem člověka.*

(25) rus. *Все дети*[17] *любят сказки.*
č. *Všechny děti mají rády pohádky.*

Obecně existenciální jmenné fráze odkazují k některé (možná také neomezené) části dané třídy. Srov. např. *někteří logici* v (26):

(26) rus. *Некоторые логики разбираются в лингвистике.*
č. *Někteří logici se vyznávají v lingvistice.*

Při anaforickém opakování anaforická jmenná fráze bude mít podle Šmeleva již obecně generickou referenci, jako např. NP *tito logici* v (27):

(27) rus. *Некоторые логики разбираются в лингвистике. Эти логики обладают хорошим языковым чутьем.*
č. *Někteří logici se vyznávají v lingvistice. Tito logici mají dobrý jazykový cit.*

Na rozdíl od přístupu Padučevové, která na generické jmenné skupiny vždy nahlíží jako na nespecifickou referenci, podle Šmeleva všechny tři skupiny – generické, abstraktně-individuální a individuální jmenné fráze – mohou mít specifickou i nespecifickou referenci.

V rámci neokauzální teorie reference zavádí Šmelev pojem *indexní reference* (*индексальная референция*). V případě indexní reference typ reference je určen typem zájmena, které se při této NP vždy používá, přičemž deskriptivní význam taková jmenná fráze buď vůbec nemá (jako např. osobní zájmena), nebo tento význam není relevantní pro identifikaci daného referentu (jako např. v případech (28)–(29), kde deskriptivní význam nestačí pro identifikaci).

(28) *ru. Какой-то дурак все испортил.*
cz. Nějaký blbec všechno zkazil.

(29) *ru. Дай-ка мне эту штуку.*
cz. Dej mi tu věc.

[17] V klasifikaci Padučevové by daná NP měla univerzální referenční charakteristiku.

II.2.2. Česká teorie reference

Z české jazykovědy je s naším tématem příbuzných následujících pět tematických okruhů:

- teorie reference a anafory (Palek 1968, Palek 1988, Hlavsa 1972, Hlavsa 1975);
- vyjadřování určenosti a kontextové zapojenosti (Adamec 1980, Daneš 1999, Hlavsa 1972, Uhlířová 1996 aj.);
- realizace anaforického vztahu v češtině (Zimová 1994);
- pravidla použití ukazovacích zájmen (Adamec 1983, Bogoczová 2000, Křížková 1971, Mathesius 1926, Nedoluzhko 2005, 2006, Slezáková 1999, Štícha 1999, Zubatý 1917 aj.);
- textová syntax a s ní související otázky koherence a koreference (Adamec 1988, Hrbáček 1994).

Českou tradici teorie reference a anafory představují především práce B. Pálka, Z. Hlavsy a P. Adamce.

B. Palek (1968) určuje místo odkazovacích vztahů v popisu jazykového systému, zachycuje rozdíl mezi denotátem a objektem skutečnosti a zakládá teoretickou bázi pro výzkum a klasifikaci odkazovacích vztahů v textu. Klasifikace denotačních statusů[18] podle Palka (1968) vychází z logických prací Quine (1960), Russela (1919), aj. a rozlišuje použití generická, singulativní (vlastní názvy, kontextově nebo situačně jedinečné objekty) a deskripce. Klasifikace je však představena pouze stručně a popis anaforických vztahů na ni nenavazuje.

Pro účely analýzy odkazovacích mechanismů Palek vyčleňuje odkazovací vztahy čtyř typů, na jejichž základě mluví o determinaci (1 a 3) a diferenciaci (2 a 4):

1. jeden objekt – jedno pojmenování;
2. různé objekty – různá pojmenování;
3. jeden objekt – různá pojmenování;
4. různé objekty – jedno pojmenování.

Na proces textového odkazování se nahlíží z dvou hledisek – referenčního a syntaktického. Na referenční úrovni se rozlišuje identifikace (identita objektů) a diferenciace (různé objekty). Na syntaktické úrovni se hovoří o odkazování (*cross-reference*) a alternaci, přičemž syntaktický proces odkazování je možný pouze v případě identity referentů antecedentu a anaforického pojmenování. Dále Palek vytváří formální model struktury textu z hlediska mechanizmů odkazování, kde se na první výskyt jmenné fráze N (*modifier*) navazují odkazy ve

[18] U Palka *classification of naming units*, viz Palek 1968, s. 55.

formě [Identifikátor[19] + ... + (pojmenování)[20]]. V Palkově modelu se však počítá pouze s jedním typem odkazování, tj. s koreferenční pronominalizací (typ *Petr – bez něj*) nebo opakováním antecedentní NP s aktualizátorem (typ *dívka – ta dívka*). Nebere se ohled na opakování jmenné fráze bez aktualizátoru (typ *Petr – Petr*), na odkazování pomocí jiného pojmenování (typ *Pálek – autor*) nebo na nekoreferenční vztahy a asociační anaforu (typ *pokoj – strop*).[21]

Podobně jako v (Palek 1968), Palkova studie *Referenční mechanismy výstavby textu* (Palek 1988) je zaměřena na analýzu odkazovacích prostředků v češtině. Autorovi nejde o referenci konkrétní jmenné fráze v textu, ale pouze o vztah mezi členy anaforického páru. K referenci NP jako takové se přihlíží jen velice málo, ta se analyzuje vždy v kontextu anafory, přičemž za anaforu se považují jenom případy, kdy anaforický člen obsahuje příslušný identifikátor.

Jmennou frází Palek (1988) rozumí spojení komponentu nominálního, který představuje obecný výraz, a komponentu instauračního, který stanoví, že jmenná fráze odkazuje na nějaký referent, tj. realizuje referenci daného nominálního komponentu. Prostředky vyjádření denotace jsou v přirozeném jazyce výrazy, které Palek nazývá instaurátory. Pojem instaurátor je podobný pojmu aktualizátor ve smyslu Padučevové, je však o něco dále rozvit tím, že v roli morfematického akturalizátoru vystupují např. osobní koncovky sloves a číslo substantiv, syntaktická pozice výrazu ve větě (např. subjektová) atd.

Při stanovení vztahů mezi denotáty v anaforických řetězcích Palek zmiňuje identitu, inkluzi (referent antecedentu v sobě zahrnuje referent postcedentu), členství a rozdíl (pak se používají alternátory) avšak podrobnější klasifikace se v práci neprovádí. Co se týče přímé analýzy možných typů reference, ta se neprovádí systematicky. Rozlišuje se mezi denotáty-konstanty (pokud situace nebo jev jsou časoprostorově určeny) a proměnnými („situace, která je pouze součástí intence mluvčího“) avšak vzápětí se tvrdí, že „konstantnost/proměnnost subjektu/objektu se vzájemně neovlivňují, nekladou žádná sémantická omezení na věty a týkají se pouze sémantické interpretace anaforických vztahů“.[22]

Pro analýzu referenční struktury textu jsou navrženy dva principy: analytický a syntetický. V rámci analytického principu se pomocí složitého systému založeného na pojmech matematické logiky a relací identity (*eq*) – různosti (*neg*), inkluze (*inc*) – neinkluze (*ninc*), disjunkce (*dis*) – nedisjunkce (*ndis*) a členství (*meb*) – nečlenství (*nmeb*) provádí analýza

[19] Identifikátor = určitý člen, zájmeno apod.

[20] Závorka znamená neobligatornost pojmenování.

[21] Tento poslední rys kritizuje také Hlavsa (Hlavsa 1972).

[22] Viz Palek 1988, s. 78.

vztahů mezi denotačními frázemi. Výsledkem jsou anaforické textové vzorce a pokus o vymezení pojmu text. Syntetický model představuje generování textové jednotky a možných referenčních vztahů v ní. Oba modely se navzájem doplňují a předpokládají.

Z Palkovy klasifikace instaurátorů vyplývá, že vztahy exoforické, endoforické (katafora a anafora), alternace a negace jsou vymezeny jenom pro referenční pozice, tj. neurčují se pro vztahy predikační povahy ani pro apozici. Exoforický odkaz se vyznačuje tím, že jeho antecedent není přímo zmíněn v předcházejícím textu, ale vyrozumívá se ze situace, zatímco při vztahu endoforickém, antecedent a postcedent „jsou v textu explicitně vyjádřeny a jim odpovídajícím denotačním vztahem je identita nebo incidence denotátů".[23] Palek dále tvrdí, že v případě exoforického vztahu je obligatorní použití identifikátoru – buď samostatného (v podobě substantivní nebo adverbiální) nebo nesamostatného v podobě adjektivní. Toto tvrzení je však poměrně diskutabilní (viz příklady v III.7.1.).

Z. Hlavsa (Hlavsa 1975) se zaměřuje na popis odkazovacích prostředků češtiny s ohledem na referenční vlastnosti a určenost jmenných frází v anaforické pozici. Podle Hlavsy jmenná fráze má designaci (vztah ke všem možným objektům vyhovujícím jejím významům) a denotaci (vztah k objektu, o kterém mluvčí vypovídá), kterou nabývá ve větě pomocí operátoru *Ref.* Popis jedinečné určenosti (determinace, operátor *Dun*) a jejich vyjadřovacích prostředků v češtině je proveden na materiálu jmenných frází s implicitně jedinečným významem,[24] jmenných frází s relační nebo kontextovou jedinečnosti (*president, ředitel, matka, mistr světa ve skoku na lyžích, dnešek, okolí* apod.)[25] a určitých deskripcí[26](typ *bratrova dcera* v případě, že mluvčí má jenom jednoho bratra a ten má jenom jednu dceru).

Klasifikace referenčních typů zaměřená na komparativní analýzu použití ukazovacích zájmen v češtině a ruštině je představena v článcích **P. Adamce** (např. v Adamec 1980, Adamec 1984 aj.). Adamec mluví o následujících typech reference:

---

[23] Ibid. 47.

[24] K nim patří "vlastní jména , osoby nebo jména jako osoby pojaté, zvířata, neživé objekty, jevy prostorové a časové; látky, jejichž jedinečnost je vymezena definičně, technologicky nebo výrobní značkou, jedinečné jevy hromadné a některá abstrakta (pojmenování vědeckých a zájmových oborů, sportu a her a negace vlastností nebo vlastnosti antonymního typu: *fátum, lucifer, hydrosféra, jih, středověk, kapron, lidstvo, čeština, akvaristika, dospělost, ilegalita, nuda, odvaha, panenství* apod" (Hlavsa 1975, s. 53.).

[25] "... Jedinečná determinace pro omezený okruh mluvčích, v omezeném časovém období nebo tím, že jsou jeho součásti". Viz Hlavsa (1975, s. 54).

[26] " ... Složený název, u něhož věcný význam sám o sobě neukazuje, že pojmenovává jednotkovou třídu, ale existuje povědomí, že není víc, než jeden objekt takovému popisu vyhovující", sem patří také superlativa a NP s pořadovými číslovkami. Viz Hlavsa (1975, s. 54n.).

- generický (*Je třeba naučit děti milovat knihu/knihy; Kolegové si musejí navzájem pomáhat*),
- podmíněně singulativní[27] (*Půjč mi nějakou knihu; Poprosím o to některého kolegu*). Daný typ je vždy neurčený,
- singulativní
  i. neurčený:
    a) neurčený pro mluvčího a adresáta (*Na chodbě jsem potkal jakéhosi pána*),
    b) neurčený jenom pro adresáta (*Mluvil jsem o tom s jedním lékařem*);
  ii. určený:
    c) určený „zvnějšku", tj. kontextem, konsituací, deixí (*Do pokoje vstoupil jakýsi chlapec s hezkou dívkou; Chlapec mi byl neznámý, ale tu dívku jsem už někde viděl. Prosím tě, podej mi ze stolu ty papíry* aj.),
    d) určený „zevnitř" – dostatečně jednoznačným pojmenováním (*Do Brna s námi pojede Mírek. Nejnovější knihu od Párala jsem ještě nečetl. Tvoje nejmladší sestra je hodně podobná na maminku*).

Problémy vznikají při pokusu o zařazení do klasifikace jmenných frází v rématu, v nichž substantiva stojící v rematické části mají pouze signifikativní funkci a jejich konkrétně předmětový charakter, a tedy i ta či ona referenční platnost jsou jakoby zatlačovány do pozadí, pociťovány jako irelevantní. Srov. (30)–(32):[28]

(30) *Ten dopis psalo dítě.*

(31) *Pojedeme tam autobusem.*

(32) *Dědeček čte noviny.*

Referenci jmenných frází v této pozici se dá interpretovat buď jako generické (Např. (30) lze přeformulovat jako (30)' *Ten dopis napsal někdo a ten někdo je dítě.*) anebo jako neurčité singulativní, protože jde o konkrétní události, vztažené ke konkrétním situacím. Adamec se přiklání k druhé možnosti.[29]

[27] *Singulativní* = specifická reference; v klasifikace Padučevové (1985) konkrétně referenční.

[28] Příklady (30)–(32) jsou převzaty z (Adamec 1980).

[29] Viz Adamec 1980, s. 159.

II.2.3. Teorie reference v pracích německojazyčných slavistů

Rozsáhlé a velice detailní dílo **T. Bergera** (Berger 1993) představuje komplexní a téměř vyčerpávající popis systému ukazovacích zájmen v současné češtině. Popis je založen na teoriích reference a anafory a je bohatě podepřen literaturou vztahující se k těmto tématům. V teoretické části práce Berger podává definice exoforického a anaforického odkazování,[30] představuje vlastní klasifikaci typů reference[31] a analyzuje souvislosti anafory a koreference.[32]

Podle Bergera jazykový výraz odkazuje *exoforicky* v případě když označovaný referent může být identifikován pouze na základě znalosti mimojazykové situace; při *endoforickém* odkazování referent jazykového výrazu může být identifikován bez znalosti mimojazykové situace, na základě předchozího nebo následujícího kontextu.

Exoforické odkazování (nebo deiktické v případě přítomnosti objektu reference v situaci promluvy) se dále klasifikuje na

- místní deixi (*Lokaldeixis*) – *V téhle knize je všechno, co potřebuješ.* (VL)
- časovou deixi (*Temporaldeixis*) – *V této době nikdy nevíš, co na Tebe čeká za rohem.* (VL)
- osobní deixi (*Personaldeixis*) – odkaz na osobu přítomnou v situaci promluvy.

Endoforické odkazování může být holoforické (na celý text) a ohraničené, které se dále klasifikuje na

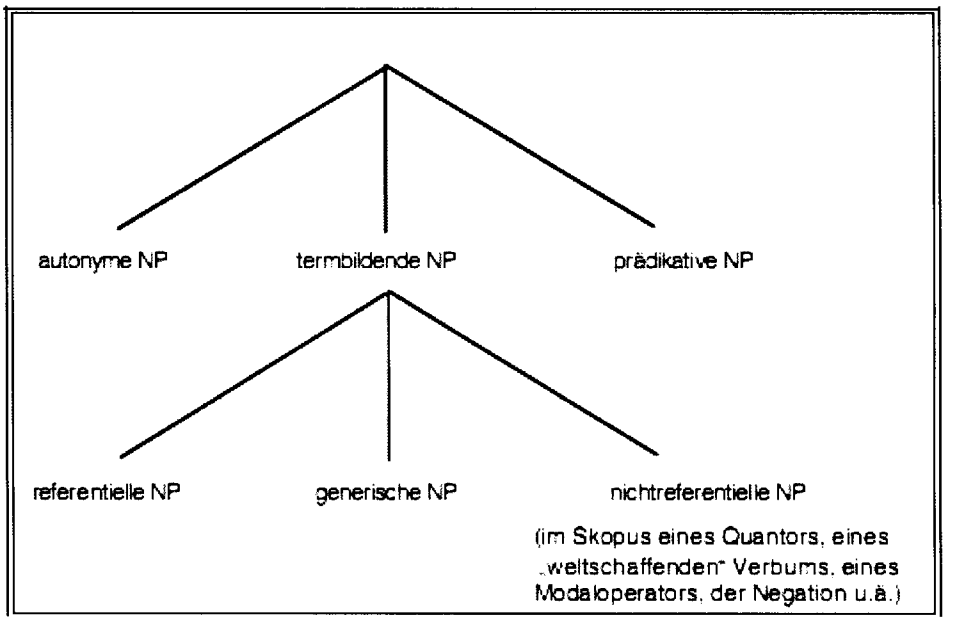
- anaforické – odkaz na segment textu v předchozím kontextu,
- periforické – oboustranné odkazování,
- kataforické – odkaz na segment textu v následujícím kontextu.

Klasifikace typů reference podle Bergera je založena na klasifikaci referenčních statusů Padučevové (1985). Rozdíly se týkají především interpretace generických jmenných frází. Padučevová (1985) zařazuje generické jmenné fráze mezi nereferenční (viz výše v této kapitole), zatímco Berger je rozebírá zvlášť a vyčleňuje je do samostatné referenční skupiny. Srov. obrázek č. 1 referenčních typů podle Bergera (1993):

[30] Viz Berger 1993, s. 227.

[31] Ibid. 291.

[32] Ibid. 298.



Obrázek č. 1: Klasifikace referenčních typů podle Bergera

Definice anafory vychází z pojmů koreference a opakování pojmenování. Podle Bergera se můžeme uvažovat o explicitní anafoře od B k A, pokud je splněna aspoň jedna z následujících podmínek:

- A a B jsou koreferenční;
- A a B obsahují stejný lexém a liší se pouze v referenčním typu (např. *létadlo – jiné létadlo*);
- existuje jmenná fráze C, která splňuje jednu z následujících podmínek:
  i. A a C jsou koreferenční, B a C obsahují stejný lexém a liší se pouze typem reference **nebo**
  ii. B a C jsou koreferenční, A a C obsahují stejný lexém a liší se pouze typem reference.

O implicitní anafoře Berger mluví v případě, že antecedent je vyjádřen formou propozice.

Kromě toho, Berger zavádí termín tzv. pseudoanafory, což je podle něj jmenná fráze, která se formálně projevuje jako anafora, nemá však v předcházejícím kontextu antecedent a pro její interpretaci není zapotřebí zapojovat mimojazykové znalosti.

Bergerova interpretace anaforického vztahu je velice široká – zahrnuje nejenom všechny případy koreference, ale rovněž opakování stejného lexému bez koreference. Tak široký přístup je však pro účely jeho práce (popisu funkcí ukazovacích zájmen) připustitelný, protože existence zájmen v analyzovaných příkladech ho neodvádí daleko od skutečné anafory. Pro

zpracování koreferenčních vztahů v našem projektu je existence takto širokého pojetí anafory velice lákavá v tom, že můžeme tvrdit, že koreferenční vztahy, které v dané etapě anotujeme na Pražském závislostním korpusu, jsou také v jistém (velice širokém) smyslu anaforické.

Habilitační práce **Imke Mendozové** (Mendozová 2004) si klade za cíl popsat systém jmenné determinace a jejích prostředků v současné polštině. Avšak k tomu je zapotřebí mít přesné referenční nástroje. Proto autorka popisuje a srovnává některé existující teorie a nakonec vytváří vlastní přehled referenčních typů, který se nám zdá velice přesný a zajímavý. Především o její klasifikaci se opíráme při zpracování vlastního schématu anotace rozšířené koreference.

Referenční status jmenné fráze se podle Mendozové (2004) skládá z jejích referenčních a textově pragmatických vlastností. Referenčními vlastnostmi jsou referenční status jmenné fráze a referenční prostor (svět), ve kterém je definován. Textově pragmatické vlastnosti vypovídají o určenosti a jedinečnosti referentu jmenné fráze v daném referenčním prostoru[33]. Autorka tvrdí, že tyto dvě roviny jsou mezi sebou neoddělitelně propojené a textová určenost je často závislá na typu reference dané NP, nelze je tedy zkoumat zvlášť.

Z referenčního hlediska je možné jmenné fráze rozdělit na referenční a nereferenční. Za nereferenční užití se považují jmenné fráze v predikativní funkci (*Marek ist Lehrer*) nebo v apozici (*Jan, ein guter Schüler, verstand sofort alles.*). Nereferenční NP se vyznačují tím, že nemohou vystupovat jako antecedenty anaforického vztahu. Na referenční rovině rozlišuje autorka následující typy reference:

- distributivní a kolektivní, kde se odkazuje k celé počitatelné a uzavřené množině objektů ($\forall$ (y) $\exists$(x) G(x,y) a $\exists$ (x) $\forall$ (y) G(x,y)) Srov. (33):[34]

(33) něm. *Die Elefanten sterben aus.*
č. *Sloni vymírají.*

- reference na otevřené množiny nebo třídy (*Referenz auf Klassen*), která se dále člení na univerzální ($\forall$x (K(x) $\rightarrow$ S(x)): srov. (34)), generickou (srov. (35)) a existenciální typy reference.

[33] Viz Mendozová 2004, s. 69.

[34] Příklady (29)–(41) jsou převzaty z Mendozové (2004).

(34) něm. *Alle Kinder essen gerne Schokolade.*
č. *Všechny děti rady jedí čokoládu.*

(35) něm. *Das Auto ist des Deutschen liebstes Kind.*
č. *Pro Němce auto je nejoblíbenější dítě.*

- Generická reference se považuje za typ zvláště komplikovaný a nejednotný, proto ji autorka klasifikuje na několik dalších podtypů (různé podtypy se mohou v jazyce vyjadřovat různými způsoby). Tyto podtypy jsou:

  i. generická reference na typ (*typen-generische*), tj. odkaz na prototypického představitele dané třídy. Srov. např. (36):

(36) *rus. Ты наивный, как ребенок.*
č. Jsi naivní jako dítě.

  ii. generická reference na třídu (*klassen-generische*), tj. odkaz na libovolný prvek dané třídy (37):

(37) angl. *The members of this club do not drink whisky.*[35]
č. *Členové tohoto klubu nepijí whisky.*

  iii. reprezentativní generická reference (*repräsentativ-generische*), která vystupuje v nezobecňujících výpovědích:

(38) něm. *Die Amerikaner landeten 1969 auf dem Mond.*
č. *V roce 1969 přistáli Američané na měsíci.*

- reference na prvek, která může být:
  i. specifická (jmenná fráze má konkrétní denotát v popisovaném světě) a
  ii. nespecifická (jmenná fráze nemá konkrétní denotát v popisovaném světě). Kontexty pro nespecifickou referenci jsou např. nereální kontexty, otázky, kondicionál, negace apod.

[35] Příklad pochází z Dahle (1975), cit. podle Mendozové (2004, s. 88).

Autorka zvláště upozorňuje na rozdíl mezi referencí na objekt a tzv. referencí na funkci (*Referenz auf Rollen*), která v některých kontextech může být i syntakticky relevantní. Za referenci na funkci autorka považuje takové užití jmenné fráze, které se tradičně (Padučevová, Donnellan) považuje za tzv. atributivní určité deskripce, srov. NP *zloděj* v (39)[36] apod.:

(39) rus. *Он наказал вора.*
č. *Potrestal zloděje.*

Anaforický vztah Mendozová (2004) chápe jako vztah definovaný na určitých jmenných frázích, který spolu se vztahem kataforickým vyjadřuje tzv. textovou určenost. Za prototypickou anaforu autorka považuje odkaz na koreferenční antecedent v předcházejícím kontextu:

(40) něm. *Otto kam in die Küche. Er öffnete den Kühlschrank und nahm sich ein Bier.*
č. *Otto vstoupil do kuchyně. (On) otevřel ledničku a vzal si jedno pivo.*

Je však možný i anaforický vztah mezi nekoreferenčními objekty, pokud je spojuje explicitní anaforický konektor. Srov.:

(41) *If John wants a hamburger, he will order it.* (z Fauconniera 1985)

II.2.4. Predikace vs. identifikace jmenné fráze v pozici přísudku

Na závěr našeho teoretického přehledu z oblasti teorie reference a anafory se zaměříme na jeden specifický problém, se kterým se setkáváme při určování referenční platnosti jmenných frází – na rozlišovaní predikačních a identifikačních konstrukcí ve větách s jmenným přísudkem. Daná problematika sice není zohledněna při zpracování koreference ve stávající anotaci PDT 2.0, je však částečně vyvoditelná ze struktury tektogramatického stromu (viz III.1.4.) a potenciálně může být pro češtinu na oanotovaných stromech vyřešena.

S problémem rozlišení identifikačních a predikačních konstrukcí se setkáváme v případě, že je daná jmenná fráze ve větě v pozici přísudku. Srov. typické příklady:

predikace (přisuzovaní kvality):	*Petr je zedník.*
identifikace referentů subjektu a jmenné části přísudku:	*Petr je právě ten zedník, který nám dělal koupelnu.*

[36] Padučevová 1985, cit. podle Mendozové 2004.

Analýza většího korpusu příkladů však ukazuje, že rozlišení mezi predikačním a identifikačním významem přísudkové jmenné fráze není vždy jednoznačné a vyžaduje hlubší analýzu. V následující kapitole se pokusíme na základě existující literatury k danému tématu najít teoretické vymezení těchto dvou typů konstrukcí.

Názory badatelů na hranici mezi predikací a identifikací se značně rozcházejí. Existuje širší a užší pojetí identifikace. Nejširší pojetí identifikace najdeme u **E. Lavricové,**[37] která za predikaci považuje pouze takové konstrukce, ve kterých není možné užít určitého nebo neurčitého členu,[38] jako např. NP *Determinantensemantikerin* v (42):

(42) něm. *Ich bin Determinantensemantikerin.*

Přísudková substantiva s neurčitým členem jsou považována za referenční s neurčitou referencí. Jako identifikační konstrukce se hodnotí i věty s generickou jmennou frází v přísudku typu *ein Insekt* v (43). V tom případě se identifikace provádí mezi třídami objektů.

(43) něm. *Die Heuschrecke ist ein Insekt.*

Takové široké pojetí identifikace kritizuje I. Mendozová.[39] Namítá, že v případě zařazení vět s neurčitým členem v přísudku mezi identifikační struktury nastává situace, kdy se jako predikace interpretují věty typu *Marek ist Lehrer* a *Johann ist gut* a jako identifikace s neurčitým objektem věty typu *Johann ist ein guter Mensch* a *Marek ist ein guter Lehrer.* Z toho, že německá věta bez členu **Marek ist guter Lehrer* je agramatická, plyne, že není možné vyjádřit predikační význam pomocí jmenné fráze, která obsahuje přídavné jméno, což není zcela logické. Kromě toho, orientace definice predikačních NP na existenci určitého nebo neurčitého členu ve větě je příliš zaměřena na konkrétní jazyk. V angličtině bude situace poměrně odlišná, srov. agramatičnost angl. **He is teacher.*

Odlišný názor nacházíme v pracích lingvistů moskevské sémantické školy (Padučevová 1987, Šmelev 1996, Aruťunovová 1976 aj.) Tady se identifikace (*высказывание идентификации, предложения тождества*) chápe také poměrně široce, avšak rozdíl mezi identifikací a predikací se provádí na úrovni sémantické za pomoci některých lexikálně syntaktických kritérií.

---

[37] Viz Lavricová 2001, s.63, cit. podle Mendozové (2004).

[38] Práce Lavricové je založena na materiálu němčiny, jde tedy o pravidla použití členu pouze v němčině.

[39] Viz Mendozová 2004, s. 73n.

N. D. Aruťunovová (1976) analyzuje sémantickou a referenční strukturu relace identifikace. Vyčleňuje několik typů situací identifikace, např. následující situace:

- situaci „detektivního vyhledávání“:

(44) rus. *Убийца старухи есть Раскольников.*
č. *Vrah(em) je Raskolnikov.*

- situaci uskutečnění snu:

(45) rus. *Это как раз и есть то, что нам нужно.*
č. *To je právě to, co potřebujeme.*

(46) rus. *Вы и есть нужный человек.*
č. *Vy jste právě ten správný člověk.*

- situaci „poznávání“:

(47) rus. *Иван Иваныч! Да ведь это ты! ты! ты!*
č. *Ivane Ivanyči! Vždyť jsi to Ty! Ty! Ty!*

- a několik dalších.

Situace identifikace se vyjadřuje v indoevropských jazycích pomocí tzv. identifikačních konstrukcí. V rámcích identifikačních vět se rozlišují nominativní (kódová) totožnost, v případě když se deklaruje stejná schopnost k referenci dvou různých nominací (srov. (48)) a tzv. denotativní totožnost, kde se deklaruje totožnost objektu se sebou samým (srov. (49)). Od obou komponentů identifikační konstrukce se vyžaduje specifická reference.[40]

(48) rus. *Цицерон есть Туллий.*
č. *Cicero je Tullius.*

(49) rus. *Бьюсь об заклад, если это не тот самый сорванец, который увязался за нами на мосту.*
č. *Vsadím se, že to je ten samý uličník, který se na nás pověsil na mostě.*

[40] Viz Aruťunovová 1976, s. 292.

Za identifikační se rovněž považují všechny výpovědi, které mají vlastní jméno v rématu, protože mohou být transformovány na konstrukci typu N1– je – N1. Srov. např. (50):

(50) rus. *Во время катастрофы пострадал Иванов.*

--> *Пострадавший - Иванов .*

č. *Během katastrofy byl postižen Ivanov.*

--> *Postižený – Ivanov.*

Aruťunovová nabízí následující testy na rozlišování klasifikačního a identifikačního vztahu[41]:

1. Možnost použití určitého determinátoru s oběma komponenty identifikační konstrukce, zatímco v klasifikačních (podle Aruťunovové tzv. inkluzívních) větách druhý komponent buď nemá žádný identifikátor, nebo má neurčitý člen. Srov. angl. *Peter is a writer* (predikace) vs. *Peter is the author of this novel.* (identifikace).
2. V identifikačních větách oba komponenty mohou být zastoupeny zájmenem, srov. např. (51):

(51) rus. *Этот молодой корнет и есть девица Дурова → Он и есть она.*[42]

č. *Tento mladý kornet skutečně je slečna Durova. → On právě je ona.*

3. Test pokračování textu: pokud zkoumaná konstrukce je klasifikační, její druhý komponent nemůže vystupovat jako subjekt následující výpovědí, protože nemá vlastní referenci, zatímco v identifikační výpovědi to možné je. Srov. např. identifikační větu (52), pro kterou pokračování typu (53) zní celkem přirozeně s oběma podměty. Avšak pokud v (52) změníme pořádek členu na (54), jde již o klasifikační větu a subjekt pokračovací věty (55) nebude koreferenční se subjektem první věty.

(52) rus. *Мой учитель – Джонс.*

č. *Můj učitel je Johns.*

(53) rus. *Мой учитель (Джонс) преподает мне математику.*

[41] Testy jsou zpracovány na materiálu ruštiny, ale jsou založeny na logicko-sémantických vztazích, jsou proto zčásti aplikovatelné i na češtinu.

[42] Ibid. 311.

č. *Můj učitel (Johns) mě učí matematice.*

(54) rus. *Джонс – мой учитель.*
č. *Johns je můj učitel.*

(55) rus. *Мой учитель преподает мне математику.*
č. *Můj učitel mě učí matematice.*

4. Test negace: pokud na identifikační větu naložíme negaci, zápor se bude vztahovat pouze na tvrzení o tom, že referenty obou komponentů věty jsou identické. Presupozice jejich existence zůstane nezměněná. Srov. např. tvrzení (57) (negace naložena na identifikační větu (56)), kde se nepopírá existence ani Napoleona ani vítěze Borodinského boje, ale pouze totožnost referentů těchto dvou výrazů.

(56) rus. *Победитель Бородинского сражения – Наполеон.*
č. *Vítězem bitvy u Borodina je Napoleon.*

(57) rus. *Ложно, что в Бородинском сражении победителем был Наполеон.*
č. *Není pravda, že vítězem bitvy u Borodina je Napoleon.*

5. Vratnost (reverzibilita) komponentů identifikačních konstrukcí typu *Venuše je Jitřenka* ↔ *Jitřenka je Venuše.*

Pokus najít formální rozdíl mezi identifikací a predikací v nemožnosti použití v identifikačních konstrukcích přísudkového instrumentálu se však Aruťunovové nepodařil. S instrumentálem se totiž používají i identifikační konstrukce (srov. (59) vedle (58)). Podle Aruťunovové je však možnost instrumentálového přísudku sekundární a je podmíněna kontaminací slovesného rámce slovesa *оказаться* (cz. ≈ *stát se*).[43]

(58) rus. *Этот клоун был мой сослуживец.*
č. *Tento klaun byl můj spolupracovník.*

(59) rus. *Этим клоуном был мой сослуживец.*
č. *Tímto klaunem byl můj spolupracovník.*

[43] Ibid. 325.

E. V. Padučevová (Padučevová 1987) v popisu identifikačních vět vychází z identičnosti referentů komponentů konstrukce. Důležité přitom je, aby oba komponenty měly specifickou (konkrétní) referenci.[44] Teoreticky by bylo možné za identifikaci považovat i konstrukce, ve kterých vystupují generické NP (srov. např. (61)), ale právě tím, že dané NP nemají specifickou referenci, se z klasifikace vylučují.[45]

(60) rus. *В маленьком населенном пункте главный праздник – ярмарка.*
č. *V malé obci je hlavním svátkem trh.*

Za identifikační se dále nepovažují následující typy vět:

1. Různá metajazyková užití – např. vysvětlení obecných pojmů (61), vnitrojazykové překlady (62) apod.[46]

(61) rus. *Аксиома – это истина, не требующая доказательств.*
č. *Axióm je skutečnost, která nepotřebuje důkazy.*

(62) rus. *Октаэдр – это восьмигранник.*
č. *Oktaedr je osmistěn.*

2. Taxonomická identifikace – srov (63)–(64) apod. Za pravé identifikační konstrukce se považuje jenom identifikace substancí.

(63) rus. *Первый урок была история.*
č. *První hodina byl dějepis.*

(64) rus. *Президент Филиппин – женщина.*
č. *Prezident Filipín je žena.*

3. Identita s abstraktními pojmy, neboť se u nich špatně určuje referent. Události a procesy mohou rovněž být v identifikačním vztahu, ale nestává se to příliš často.

4. Věty vyjadřující metaforickou identitu, např. (65):

[44] Srov. klasifikaci denotačních statusů v Padučevová 1985, a Aruťunovová 1976.

[45] Viz Padučevová (1987, s. 154), ale také Weiss (1978, s. 245).

[46] Příklady (60)–(70) jsou převzaty z Padučevové (1987).

(65) rus. *Государство – это я.*
č. *Stát jsem já.*

Z hlediska aktuálního členění Padučevová upozorňuje na zdánlivou subjektivnost slovosledu v identifikačních větách. Říkáme-li větu (66), o Venuši se předpokládá, že adresát je schopen ji identifikovat, zatímco pro Jitřenku je zapotřebí další vysvětlení. Tím se však jedině znázorňuje fakt, že k tomu, aby určitý výraz byl v tématu, je dostatečná kontextová zapojenost a není nutná souvislost s aktuálními znalostmi adresáta o světě.

(66) rus. *Утренняя звезда – это Венера.*
č. *Jitřenka je Venuše.*

Na sémantické úrovni je podle Padučevové (1987) možné hovořit o čtyřech typech sémantické identifikace:

1. První komponent je atributivně použitá deskripce, jako v (67):

(67) rus. *Столица Перу – Лима.*
č. *Hlavním městem Peru je Lima.*

2. Provádí se shoda mezi lokálně a temporálně identifikovaným objektem a jeho obecným názvem (68) nebo se deskriptivní znalost objektu identifikuje se samým objektem (69):

(68) rus. *Эта освещенная магистраль – улица Кропоткина.*
č. *Tato rozsvícená třída je ulice Kropotkinova.*

(69) rus. *Маяковский – это я.*
č. *Majakovský jsem já.*

3. Oba komponenty jsou obecná jména a „globální názvy", adresátovi se „nabízí možnost změnit model světa, ve které odpovídající jména označují dva různé objekty na takový, kde označují jeden stejný objekt".[47]

[47] „Говорящий предлагает слушающему заменить модель мира, в которой соответствующие имена обозначают два разных объекта, на такую, в которой они обозначают один и тот же объект, ср. *Утренняя звезда и Венера – это одно и то же небесное тело* ". Viz Padučevová (1987, s. 161).

4. Komponenty vztahu jsou dvě různé *kvalifikace* jednoho objektu. Identifikační vztah je spojuje:[48]

(70) rus. *Хозяин гостиницы – оценщик в городском ломбарде.*
č. *Majitel hotelu je odhadce v městské zastavárně.*

Padučevová (1987) naznačuje, že rozdíly mezi uvedenými sémantickými typy identifikace se projevují také v syntaxi (jako např. různé možnosti atributivních a komunikativních modelů, možnosti ne/použit výrazu *это* (*to*) jako součást druhého komponentu konstrukce, různá pravidla užití zájmen v prvním komponentu), v morfologii (aplikace na časový systém ruštiny) apod. Tyto rozdíly se však zde dále nerozebírají.

A. D. Šmelev (Šmelev 1996) tvrdí, že cíl identifikační výpovědi je poskytnout adresátovi výpovědi možnost přesně lokalizovat referent prvního komponentu.[49] Podle Šmeleva[50] identifikační výpovědi mohou být homonymní s výpověďmi s predikativní NP v přísudku, např. (71) lze pochopit na jedné straně jako charakteristiku Ivana, na druhé straně jako odpověď na otázku *Kdo je Ivan?*, což je podle Šmeleva případ tzv. vysvětlující identifikace.

(71) rus. *Иван – мой друг.*
č. *Ivan je můj kamarád.*

Tuto homonymii lze vyřešit pomocí perifráze (72) a (73):

(72) rus. *Иван – мой друг → Иван мне друг: predikace*
č. dosl. *Ivan je můj kamarád. → Ivan je mi kamarád.*

(73) rus. *Иван – мой друг → Иван – это мой друг: identifikace*
č. dosl. *Ivan je můj kamarád. → Ivan – to je můj kamarád.*

Za ukazatele identifikace lze považovat ukazovací zájmeno *eto* (*to*) v pozici prvního komponentu, a užívané zároveň s ním determinátory *etot* (*ten*), *odin* (*jeden*) jako součást druhého komponentu konstrukce, srov. *Это один мой друг* (*To je jeden můj kamarád*).

[48] „Устанавливается принадлежность двух ипостасей или срезов одному и тому же объекту". Ibid. 162.

[49] K obrácené komunikativní struktuře těchto konstrukcí viz (Weiss 1978).

[50] Šmelev 1996, s. 177.

Naopak, na charakterizační funkci jmenné fráze ukazuje podle Šmeleva osobní zájmeno ve funkci subjektu (*Он мой друг* (*On je můj kamarád*)) a některé další specifické prostředky, které zdůrazňují predikativnost druhého komponentu. Tato kritéria však neplatí vždy. Např. ve větách s *eto* (*to*) typu (74)–(75) apod., které Šmelev pokládá za zvláštní druh identifikace, kde „kvalifikace se tváří jako identifikace“,[51] je intuitivně přirozenější postulovat vztah charakterizační.[52] Podobných příkladů se však v jazyce najde mnoho, čímž se jednoznačnost souvislostí *eto* ve funkci subjektu s identifikační povahou odpovídající výpovědí výrazně zpochybňuje.

(74) rus. *Это необыкновенный ребенок.*
č. *To je neobyčejné dítě.*

(75) rus. *Это талант.*
č. *To je talent.*

D. Weiss (Weiss 1978) poukazuje na časté smíšení pojmů *Identität* (identita, totožnost, úplná shoda), ve kterých jde o totožnost referentů prvního a druhého komponenu konstrukce a *Identifikation* (identifikace, ztotožnění), které může být přítomné i ve výpovědích s predikační NP v druhém komponentu. Daný problém nastává často i pro češtinu – identifikaci je možné rozumět neterminologicky i jako zařazení objektů do skupiny, zde se však zabýváme pouze příklady označovanými Weissem jako *Identität*. Identifikační věty se podle Weisse charakterizují dvěma kriterií: koreferencí referentů obou komponentů konstrukce a zvláštní komunikativní strukturou, o které tvrdí, že oba výrazy konstrukce nejsou z komunikativního hlediska rovnocenné; zatímco referent prvního výrazu je představen mluvčím jako neznámý, referent druhého výrazu je představen jako známý.[53]

Podle Weisse identifikační věty informují posluchače, že výraz X, referent, který mu ještě není znám (aspoň podle názoru mluvčího), se dá zaměnit na výraz Y, který se pokládá za známý; dále oba výrazy mohou být používány pro označení téže skutečnosti.

---

[51] „Такое употребление означает, что объект (чаще всего – лицо) рассматривается как «персонификация» указанного качества. [...] По существу здесь характеризация маскируется под идентификацию“. Šmelev 1996, s.178.

[52] Srov. k tomu také Mendozová (2004, s. 75).

[53] „... Die beiden verglichenen Ausdrücke sind kommunikativ nicht gleichwertig, insofern als der Referent des ersten vom Sprecher als unbekannt, derjenige des zweiten als bekannt vorausgesetzt wird. (Weiss 1978, s. 228)“.

Situace nominalizace (*Benennung*) podle Weisse není identifikační. V potenciálně dvojznačných větách (např. rus. *Я Распутин* (*Já jsem Rasputín*)) identifikaci poznáme podle toho, že její druhý komponent bude prezentován posluchači jako známý, tj. takový, kterému může být jednoznačně přisouzen referent, zatímco v nominalizačních větách je situace opačná: téma může být známé a kontextově zapojené, zatímco réma nikoliv. Srov. (76):[54]

(76) rus. *Жил-был один король. Короля звали Вася.*
č. *Byl jednou jeden král. Ten se jmenoval Vasja.*

Při výkladu vztahu totožnosti Weiss používá následující kritéria, která pomáhají rozlišit identifikaci a predikaci:

a) syntaktické kritérium: důležitá je pozice ve větě. Pokud je obecné jméno v identifikační konstrukci na začátku věty v pozici podmětu, zachovává se identifikační význam. Jakmile se přemístí na pozici přísudku, tento význam se ztrácí.[55]

b) syntaktické kritérium: predikační větu můžeme parafrázovat jako běžnou predikaci: [56]

(77) rus. *Я был автором этой статьи.* → *Я написал эту статью.*
č. *Byl jsem autorem toho článku.* → *Napsal jsem ten článek.*

c) syntaktické kritérium: na rozdíl od Aruťunovové Weiss tvrdí, že určitý determinátor u druhého komponentu není jednoznačným ukazatelem identifikace. Např. (78) může být pojata dvěma způsoby: a) patří k množině lidí, kteří zabili tu starou paní; pravděpodobně tato množina obsahuje jenom ten jeden prvek. Význam věty odpovídá '(on) zabil tu starou paní'; b) je identický s tím člověkem, kterého adresát zná jako toho, kdo zabil tu starou paní. Význam věty neodpovídá '(on) zabil tu starou paní'. V prvním případě jde o predikaci, v druhém – o identifikaci.[57]

(78) rus. *Он – убийца старухи.*
něm. *Er ist der Mörder der alten Frau.*
č. ≈ *On je vrah staré paní.*

[54] Příklad viz Weiss 1978, s. 227.

[55] Ibid. 232.

[56] Ibid. 233.

[57] Ibid. 237.

d) dobré lexikální kritérium na identifikaci pro ruštinu, které funguje v případě subjektivního slovosledu[58] – možnost dodat „и есть", jako např. ve větě *Зевс – это Юпитер.*

Vztah identifikace je možný pouze pro referenční NP, nikoliv pro výpovědi typu (79)–(80), kde se postuluje vztah synonymie.

(79) rus. *Кимарить – это спать.*
č. dosl. *Klimbat je spát.*

(80) rus. *Гардероб – это то же, что платяной шкаф.*
č. *Šatna je totéž jako skříň na šaty.*

Podstatu identifikační výpovědi předvádí Weiss v tabulce se syntaktickou a referenční informací o komponentech analyzované konstrukce:[59]

první komponent identifikační konstrukce, X	**druhý komponent identifikační konstrukce, Y**
téma	réma
kontextově zapojen	není kontextově zapojen
není (dostatečně) určitý	známý, určitý

Tabulka č. 2: Charakteristiky identifikační věty podle Weisse a Paducčevové.

V češtině je problém rozlišování predikačních a identifikačních vět zmíněn u Hlavsy (Hlavsa 1975). Uvažuje se o predikačním významu NP *mechanik* v (81), ale ve větě (82) *mechanik* už není predikační, tady jde o ztotožnění, daná NP má tedy funkci denotační.

(81) *Jan je mechanik(em).*[60]

(82) *Jan je ten mechanik.*

Zdá se však, že tento problém není pro češtinu ani zdaleka vyřešen. Je možné doložit příklady, kde se s přísudkovým substantivem použije ukazovacího zájmena, ačkoliv konstrukce není jednoznačně identifikační. Existují totiž případy tzv. údajné identifikace, kdy věta vypadá

[58] Srov. k tomu Paducčevová (1987, s. 157n.), Aruťunovová (1976, s. 312).

[59] Podobnou tabulku uvádí Paducčevová (1987).

[60] Příklady (81)–(84) jsou převzaty z Hlavsy (1975).

jako identifikační, ale ve skutečnosti spíše přisuzuje vlastnost. Jsou to věty typu (83)–(84) apod.:

(83) *To byla na tom právě ta zvláštnost.*

(84) *To byl ten vtip.*

Význam (83) se nezmění, pokud větu přeformulujeme na (85), která však už není identifikační.

(85) *To bylo na tom právě zvláštní.*

Také se ukazovací zájmeno může objevit v konstrukcích s osobními zájmeny jako první komponent identifikace.[61] Srov. (86)–(88):

(86) *To jsem já. Já jsem to moře i ten muž, ten polibek vydechnutý z temného stínu úst patří mně.* (Karel Čapek, Povětroň)

(87) *Je on ten princ, který si mne odtud odvede?* (Páral, V., Profesionální žena)

(88) *Snad vy nejste ten Frič, ten Josef Frič, to jste vy?* (Macura, V., Komandant)

Předpokládáme, že analýza většího počtu konstrukcí na anotovaném korpusu může přispět k řešení dané problematiky.

II.3. Anotace – literatura z oblasti zpracování koreference v počítačové lingvistice

V dané etapě vývoje počítačové lingvistiky, zpracování vztahů přesahujících hranice jedné věty začíná být velice populární. Do oblasti textových vztahů patří Filadelfský korpus zachycující sémantické významové vztahy v diskurzu Penn Discourse Treebank (Prasad a kol. 2008), výzkum orientačních bodů (projekt MapTask), koreference, asociační anafory (bridging relations), diskurzivní deixe (odkaz na implicitní antecedent v předcházejícím textu, na

[61] Srov. k tomu Šmelev (1996, s. 177n., viz výše), který tvrdí, že podobné konstrukce nemohou být identifikační, protože použijeme-li osobního zájmena 3. osoby, znamená to, že objekt už můžeme identifikovat a další identifikace není nutná. Při použití identifikace s osobním zájmenem dochází ke konfrontací modálních rámců a tedy k anomáliím. Jisté výjimky jsou možné jenom pro osobní zájmena 1.osoby.

segment textu) a jiných prostředků koheze, které můžeme velice široce zařadit do vztahů typu anaforického.

V současné době již mnoho velkých textových korpusů má anotaci mezivětných vztahů textové koheze. Mezi nimi je i velký počet anotačních schémat koreference a asociační anafory. V následující kapitole představíme přehled některých zahraničních prací a anotačních projektů. U některých projektů uvedeme také srovnání anotačních principů a pravidel s naší anotací.

Existující anotace anaforických jevů můžeme velmi zevrubně rozdělit na dvě skupiny – anotace na základě podrobně lingvistiky propracovaného anotačního schématu a anotace s jednoduchým schématem nebo bez něj, na základě statistických dat, přičemž anotace s jednodušším schématem mívají vetší textové korpusy s hotovou anotací. K první skupině patří např. Krasavina – Chiarchos (II.3.5.), Müller – Stube (II.3.7.), MATE a MATE/GNOME (II.3.3.) a jiné. Druhá skupina zahrnuje MUC (II.3.1.), ACE (II.3.2.) a MapTask (Anderson a kol. 1991).

II.3.1. Anotační schéma MUC

Projekt MUC (Message Understanding Conferences) je zaměřen na zpracování a vylepšení metod automatické extrakce informace z textů. Jde o sérii soutěžních konferencí (MUC-1 – MUC-7), kde se na základě předepsaného formátu prováděly evaluace různých metod extrakce dat. Na MUC-6 a MUC-7 se přidal i úkol rozpoznání pojmenovaných entit a koreference. Formát zpracování dat z oblasti koreference je představen v Hirschmanovi (1997).

Ačkoliv je schéma anotace koreferenčních vztahů u Hirschmana již poměrně starší, je stále jedno z nejpopulárnějších, má širokou oblast použití ve srovnání s jinými anotačními schématy a je pravidelně ověřováno na velkých textových korpusech. Vypracované pro účely extrakce informace, anotační schéma má za úkol najít v textu a propojit co největší počet koreferenčních párů, také v případě, že neodpovídají intuitivnímu chápání koreference, ale mají podíl na kohezi textu. Nutnost zachování jednotnosti anotace a nejvyšší mezianotátorské shody (cca. 95%) a snaha o možnost co nejrychlejší (tedy i nejlevnější) anotace vedla k rozhodnutí anotovat pouze jeden koreferenční vztah IDENT (identita referentů) mezi jmennými frázemi. Koreferenčního vztahu v MUC se nemohou zúčastnit klauze, taktéž se jako koreferenční neanotují jmenné fráze, které jsou mezi sebou ve vztahu části/celku, množiny/podmnožiny apod.

Koreferenční vztahy v MUC jsou zaznamenány pomocí SGML atributu REF, který ukazuje na ID libovolného koreferujícího výrazu. Následující příklad je ukázkou tagování, v níž se tvrdí, že jmenné fráze *Lawson Mardon Group Ltd.* a *it* jsou koreferenční.

<COREF ID= „100“>Lawson Mardon Group Ltd.</COREF> said <COREF ID=„101“ TYPE= „IDENT“ REF= „100“>it</COREF> will continue the investigations.

Tabulka č. 3 představuje o něco detailnější popis anotace koreference v MUC-7:

	Realizace v MUC
typy anotovaných vztahů	pouze identita (IDENT)
co se anotuje	• substantiva, • jmenné fráze (za jmenné fráze jsou považovány také data (*January 23*[62]), peníze (*$1.2 billion*) a procenta (*17%*)), • zájmena (osobní a ukazovací, včetně posesivních).
co se neanotuje	• wh-slova, • věty, klauze (ani v případě koreference s jmennou frází), jiné slovesné formy, (gerundia, infinitivy) • elidované členy.
anotace predikace a apozice jako koreference	ano
velikost anotované jednotky	maximální jmenná fráze
pojmenované entity	• anotují se, • části pojmenovaných entit se neanotují (*Equitable of Iowa Cos. ... located in Iowa*).
adjektiva	anotují se pouze substantiva v adjektivní pozici, které jsou koreferenční s pojmenovanou entitou nebo se syntaktickou hlavou větší jmenné fráze

Tabulka č. *3: Anotační schéma v MUC (Hischman 1997)*

Pro MUC-7 Hirschman předkládá způsob řešení anotace koreference generických jmenných frází a jiných jmen s nespecifickou referencí:

[62] Data se v MUC-7 anotují vždy jako jediná atomická jednotka, čili mezi *1995* a *this year* v *January 5, 1995 – this year* koreferenční vztah nebude postulován.

„The general principle for annotating coreference is that two markables are coreferential if they both refer to sets, and the sets are identical, or they both refer to types, and the types are identical."[63]

Avšak, jak uvadí Hirschman, najde se mnoho problematických příkladů, kde není zřejmé, jde-li o odkaz na typ, nebo na množinu. Neexistuje jednoduchý algoritmus pro určování ontologické kategorie referentu, tedy anotační schéma MUC pro rozhodování používá několik heuristických pravidel, orientovaných na angličtinu. Například se tvrdí, že *most occurrences of bare plurals refer to types or kinds, not to sets*. Tedy v (1)[64] jména *producers*, *Producers* a *they* mají generickou referenci a odkazují na typ, proto jsou označena jako koreferenční, i když neplatí, že množiny výrobců jsou ve všech použitích daného jména identické.

(1) angl. *...producers don't like to see a hit wine increase in price... *Producers* have seen this market opening up and *they*'re now creating wines that appeal to these people.*
č. *Výrobce netěší, pokud roste cena populárního vína ... Výrobci si všimli, že se tento trh otvírá, a nyní proto [oni] vyrábějí vína, která těmto zákazníkům vyhovují.*

Představené řešení koreference generických jmenných frází je však ještě poměrně úzké. Jak ukážeme později (III.4.2.1.), odkazy na typy mají širokou škálu gradací, která může být velice složitá pro rozpoznání a přesnou diferenciaci a problematická z hlediska možností jejího potenciálního využití. Tento problém se u Hischmana neanalyzuje.

Anotace koreferenčního vztahu mezi subjektem a jmennou částí přísudku v přísudku jmenném se sponou vede k nutnosti řešení problému koreference jednotek měnících se v čase. Srov. (2):

(2) angl. *The temperature was 90 yesterday and has already reached 95 today. This sets a new record high.*
č. *Včera dosahovala teplota 90° F a dnes již dosáhla 95° F. To představuje nový teplotní rekord.*

[63] „Základní princip anotace koreference je takový, že dvě jednotky se považují za koreferenční, pokud obě odkazují na množiny a ty množiny jsou identické, nebo obě odkazují na typy, a ty typy jsou identické". (překlad – *AN.*)

[64] Příklady (1)–(3) jsou převzaty z Hirschman (1997).

V MUC se anotuje identická koreference mezi „temperature“ a „90“. Teplota se pak ale mění na 95, avšak nabízející se možnost anotace „temperature“ a „95“ vede k tomu, že do množiny koreferenčních elementů se dostávají „90“ a „95“, což neodpovídá skutečnosti. Proto se na základě konvence anotuje vždy jenom nejbližší koreferenční hodnota, čili v daném příkladě pouze koreference mezi „temperature“ a „90“. Hodnota „95“ se dostává do jiné koreferenční množiny: „95“ – „this“ – „a new record high“.

V případě, když jsou obě měnící se hodnoty součástí stejné klauze, se jako koreferenční se subjektem, vybírá aktuálnější hodnota. Tedy v (3) „the stock value“ je koreferenční s $9.15 nikoliv s $8.05:

(3) angl. *The stock value rose from $8.05 to $9.15.*
č. *Hodnota akcie vzrostla z $8.05 do $9.15.*

II.3.2. Anotační schéma ACE

Po roce 1999 vedoucí funkci konferencí MUC převzal program ACE (Automatic Content Extraction, Doddington a kol. 2004), zaměřený na vývoj technologií pro podporu automatického zpracování přirozeného jazyka (NLP). ACE zpřístupnil koreferenčně oanotované korpusy pro angličtinu, arabštinu, čínštinu a částečně pro španělštinu. Program ACE je zaměřen na identifikaci a správné zařazení sedmi typů entit – osoba (PER), organizace (ORG), geopolitická entita (GPE), místo (LOC), zařízení (FAC), dopravní prostředky (VEH), zbraň (WEA). Každá z těchto skupin má od tří do devíti podtypů, např. dopravní prostředky jsou dále členěny na *air, land, subarea-vehicle, underspecified, water*. Objekty, které nezapadají do této klasifikace, jsou zcela ignorovány. Kromě uvedených typů a podtypů, každá entita je zařazena do jedné z následujících tříd, označujících typ její reference:

- specifická reference (SPC) – pojmenováaní odkazující na konkrétní, specifickou a jedinečnou entitu reálného světa:

(4) angl. *A crowd of angry Muslims set fire to a hotel.*
č. *Dav rozezlených muslimů podpálil hotel.*

- generická reference (GEN) – reference na třídu nebo typ objektů,
- výraz v negativním kontextu (NEG) – reference na prázdnou množinu nebo typ objektů,

(5) angl. *No sensible lawyer would take that case.*
č. *Tohoto případu by se žádný rozumný advokát neujal.*

- nespecifická reference (USP) – reference na nevybrané objekty v modálních, budoucích, hypotetických a otázkových konstrukcích:

(6) angl. *Many people will participate in the parade.*
č. *Přehlídky se zúčastní mnoho lidí.*

Každá entita v ACE má vyplněný atribut TYPE, který ji zařazuje do množiny pojmenovaných entit (NAM), substantiv včetně tzv. bare nouns (NOM) a zájmen (PRO) včetně wh-výrazů, partitivních konstrukcí (např. *half of the team – polovina mužstva*) a NP bez syntaktické hlavy (např. *the dead – mrtvý*). Rozdíl se provádí také mezi referenčními (REF) a atributivními (ATR) použitími. Jako ATR se označují predikativní NP, NP v apozici a jmenné modifikátory, přičemž stejná entita může mít jak referenční, tak i atributivní hodnotu.

Na označování koreference jako takové se v anotačním schématu ACE neklade velký důraz. Všechny entity, které jsou rozpoznány jako koreferenční, jsou označeny tímto vztahem. Přitom není důležité, jak jsou tyto koreferenční entity označeny, jestli jsou v anaforickém vztahu a do jaké míry interpretace příp. anaforických pojmenování vyžaduje znalosti světa. Rozhodující v ACE je tedy úkol správného zařazení pojmenování do třídy a přisouzení mu správných referenčních atributů. Koreferenční propojení se pak realizuje téměř automaticky.

Zajímavým způsobem jsou vyřešeny metonymické přenosy, kdy název entity je použit pro pojmenování jiné entity, např. název hlavního města použitý pro sportovní mužstvo nebo státní vládu. V tom případě se označuje koreference podle smyslu a pojmenovací výraz má pozitivní hodnotu atributu „metonymy_mention". Také jiné metonymické přenosy, např. místo – organizace v (7) jsou označeny a zaznamenány v daném atributu.

(7) angl. *Wouters died an hour later at St. John Macomb Hospital (místo). The suspect died later the same night, hospital (organizace) spokeswoman O'Grady said Thursday.*
č. *Wouters zemřel o hodinu později v Macobské nemocnici sv. Jana. Podezřelý zemřel později téže noci, jak sdělila ve čtvrtek mluvčí nemocnice O'Grady.*

Samozřejmě omezení množství typů označovaných entit zjednodušuje anotační úkol ACE. Avšak jejich výsledky jsou velmi dobré.

II.3.3. Anotační schéma MATE a její aplikace (korpusy GNOME a VENEX)

Podrobným schématem anotace koreference s již bohatou vývojovou historií je zpracování koreferenčních a anaforických vztahů v rámcích projektu MATE. Projekt je primárně orientován na deskriptivní víceúrovňovou anotaci mluvených dialogů anglického jazyka, ale předpokladem je rovněž jednoduchá aplikace schémat na anotaci jiných druhů textů a také na jiné jazyky, než je angličtina. Koordinátorem zpracování koreference v MATE je Massimo Poesio, jemuž patří také většina prací o anotačním schématu, aplikaci a evaluaci anotace koreference v rámcích daného projektu (Poesio 2004a a další).

MATE je koncipován takovým způsobem, že anotátor si může vybrat, které elementy (zájmena, jmenné a předložkové fráze, klauzy apod.) a do jaké míry podrobnosti (pouze identita, asociační anafora a které její typy) bude anotovat. Dialogy se anotují na úrovni prozodické, morfologické, syntaktické a komunikační (dodržování Gricových maxim apod.), anotuje se rovněž rovina řečových aktů (dialogue acts) a koreference (Mengel 2000). Zvláštní pozornost je věnována problémům, které vznikají na hranici vymezovaných úrovní.

Podle Poesia anotace anaforických vztahů nemůže existovat bez přesně definovaných cílů, neexistuje tedy takový pojem jako „general-purpose anaphoric annotation" (Poesio 2004a). Nicméně, cíle anotace koreference v projektu MATE a jeho rozšířeních (především v projektu GNOME (Poesio 2004a)) jsou poměrně široké. Za zmínku stojí možnost použití korpusu pro řešení aktivovanosti (salience) (Poesio a kol. 2000b, Poesio – Nissim 2001, Poesio – Modjeska 2002, Poesio 2003, aj.), statistické modely generace přirozeného jazyka (Poesio 2000c, Henschel a kol. 2000, Cheng a kol. 2001, aj.), automatické rozřešení anafory (anaphora resolution) (Poesio – Alexandrov-Kabadjov 2004, Kabadjov a kol. 2005), automatické zpracování určitých deskripcí (Poesio – Vieira 1998, Vieira – Poesio 2000, Poesio a kol. 2004c), anotace asociační anafory (Poesio 2004a, Vierra – Poesio 2000 aj.), aplikaci a evaluaci vztahů asociační anafory (např. pomocí WordNetu v Vieira – Teufel (1997), pomocí Google a WordNetu v Poesio a kol. 2004b). V neposlední řadě je cíl zpracování bohatě anotovaného korpusu pro řešení teoretických lingvistických úloh.

Původní verze projektu MATE je pojata jako vypracování teoretického rámce pro anotaci anaforických vztahů, který zahrnuje možnost identifikace elementů textu, které se mohou zúčastnit anaforického vztahu (tzv. „markables"), a další specifikaci těchto vztahů (Poesio a kol. 1999). Vyčleňuje se základní schéma (Core Scheme) přibližně odpovídající anotačnímu schématu v MUC (viz II.3.1.) a rozšíření, které uživatelé mohou aplikovat podle vlastní potřeby. V této verzi anotačního schématu se jako základní vyčleňuje přímá anafora (identická

anafora, v terminologii Poesia *direct anaphora*) pro koreferenční vztah určitých deskripcí se stejným řídícím členem (stejnou syntaktickou hlavou). Za asociační anaforu se považuje jednak vztah mezi koreferenčními deskripcemi s různou syntaktickou hlavou *(dům – budova),* jednak vztah mezi nekoreferenčními deskripcemi (*dům – byt*) (Vieira – Poesio 2000, s. 544), přičemž anotace těchto vztahů nepatří do základního schématu a předkládá se pouze jako teoretická možnost pro budoucí anotaci.

Vlastní aplikace základního anotačního schématu je realizována na korpusu GNOME (Poesio 2000c, 2000d). Tento korpus se skládá ze tří subkorpusů (muzejní korpus – popis muzejních objektů, farmaceutický korpus a korpus didaktických dialogů z korpusu SHERLOCK) a je relativně malý (cca 1500 vět, cca 9000 NP). Na tomto korpusu byla provedena anotace několika úrovní informace (např. rozdělení na klauze, věty a výpovědi, rétorická struktura, vyčlenění jmenných frází a přidělení atributů). Vymezení těchto atributů je zaměřeno na zpracování anaforických a jiných textových vztahů, a je tedy pro náš výzkum velice důležité. Z těchto důvodů uvedeme tu anotované v GNOME atributy a podrobně je rozebereme.

- Životnost Má významy *životný*, *neživotný* a *nejde specifikovat* – především u koordinačních konstrukcí;
- Typ NP. Daný atribut určuje typ jmenné fráze, která je anotována jako „markable" – ne/ určitá, posesivní, jmenná fráze s ukazovacími zájmeny *this* a *that*, pojmenovaná entita, zájmeno apod. Zvlášť se vyčleňují jmenné fráze s významem míry (*gram, počet* apod.), číslové výrazy *(first car)*, kvantifikátory apod.
- Počitatelnost. Má významy *count-yes* pro počitatelné, *count-no* pro nepočitatelné, *undersp-count* pro nejasné případy a *no-count* pro jednotky, na kterých se daná kategorie neurčuje (např. pro koordinované struktury)
- Deiktičnost. Má významy *deix-yes* pro odkazy na předměty mimojazykové reality a pro zájmena první a druhé osoby, *meta* pro odkazy na segmenty textu (sekci, odstavec, stránku, obrázek, název apod) a *deix-no* pro ostatní výrazy.
- Generičnost informuje, referuje-li daná jmenná fráze genericky nebo specificky. Atribut GENERIC má významy *generic-no* pro jmenné fráze se specifickou referencí, odkazující na konkrétní materiální objekty, *generic-yes* pro jmenné fráze, které odkazují na typy objektů (*Tigers are dangerous animals*) a *undersp-generic* pro nejasné příklady. Význam *generic-yes* atributu GENERIC se připisuje všem predikativním NP, většině „bare nouns" (substantivum bez členu) s významem látek a chemikálií (*I like*

music / wine / bread. Estracombi TTS patches contain oestradiol and norethisterone acetate.) nebo s abstraktním významem (*change of life, scenes from mythology* apod.). Také zájmeno může mít generickou platnost (*The man who gives his paycheck to his wife is wiser than the man who gives it to his mistress*). Význam *undersp-generic* se používá pro koordinační konstrukce, kde jeden člen je *generic-no* a druhý je *generic-yes*, a v případě více možností interpretace. Při anotaci většího počtu jmenných frází nicméně poměrně často nastávají problematické případy, kde o generičnosti daného jména nelze jednoduše rozhodnout. Problematické případy se v GNOME řeší používáním veškerých konvencí (např. gerundia jsou vždy generická, vlastní jména, včetně názvů nemocí, jako epilepsie jsou vždy negenerická), které mají obecnou tendenci všechny problematické z toho hlediska jmenné fráze anotovat jako generické.

- Logický typ reference. Má následující významy: kvantifikace (*Every department has different procedures for hiring, How many books did you buy?*) zahrnující distributivní referenci, *coord* pro koordinační struktury, *pred* pro jmenné fráze, které nereferují, ale predikují, tj. vypovídají o vlastnostech objektu (např. *John is an astronomer* ale také *The egg becomes transformed into a beautiful as well as precious object* a apozice typu *Anne-Marie Shillitoe, an Edinburgh jeweller*) a *term* pro jmenné fráze odkazující na termy.
- Funkčnost jmenné fráze. Atribut má následující významy: vlastní jméno, diskurzivní funkce (všechna zájmena a určité NP s jasným antecedentem), sémantická funkce (možné druhy situačních unik, jmenná fráze skladající se z jednoho abstraktního jména a specifikující vedlejší věty (*the question whether the definite article is a numeral*), valenční substantiva, superlativa apod.), aj.
- Ontologický status. Atribut zařazuje jmennou frázi mezi abstraktní/konkrétní a má několik následujících subatributů: osoba, substance (*voda, zlato*), léky,[65] jiná konkréta, místní a časové NP, události (*in the Dutch wars of 1672 – 1678*), nemoci, abstrakta a skupina *ostatní* pro nejasné případy. Problém ontologické klasifikace vzniká u zájmen, kde se používá význam ontologického statusu plnovýznamového antecedentu. Kolektivním a generickým NP se ontologický status přiděluje podle ontologického statusu elementů, ze kterých se skládá (např. *a group of people* se anotuje jako *osoba).*
- Reference – přímá deiktická nebo vlastního jména, omezená kvantifikátorem (*Few of Carlin's wealth clientele would have put their money in this area*).
- Struktura referentu: atomický, množinový, ostatní.

[65] Tato kategorie je pravděpodobně dána specifikou jednoho z korpusů GNOME.

- Syntaktická funkce (subjekt, objekt, predikát, adjunkt aj.), gramatické číslo (singular, plural, jiné), gramatická osoba (první, druhá, třetí) a rod substantiv (s významy maskulinum, femininum, neutrum).

Srovnáme-li anotované atributy jmenných frází v GNOME s gramatickou informací tektogramatické roviny v Pražském závislostním korpusu, který je předmětem naší analýzy, vychází najevo, že atributy `gram` v PDT 2.0 obsahují sice širší spectrum gramatických údajů (Mikulová a kol. 2005), jsou však orientovány především na řešení jiných úloh, než koherence textu. Především pro výzkum anaforických vztahů v PDT chybí různé druhy sémantické a referenční informace: informace o abstraktnosti/konkrétnosti, určitosti/neurčitosti, generičnosti, termovosti/predikativnosti daného jména aj. (srov. tabulku č. 4). Z gramatických funkcí by se pro řešení otázky koreference a její následné lingvistické analýzy velice hodila rovněž informace o počitatelnosti daného jména. Některé informace, které jsou v GNOME zahrnuty jako atributy jmenné fráze, jsou v PDT 2.0 vyvoditelné ze syntaktické struktury tektogramatické roviny (např. použití jména s ukazovacím zájmenem nebo bez něj, existence kvantifikátorů nebo číslovky apod.; podobně, syntaktická struktura tektogramatické roviny a význam atributu `is_member` informuje o tom, je-li daná jmenná fráze součástí koordinační konstrukce), nebo z významu funktorů (např. jmenné fráze s významem míry mají v PDT závislý uzel s funktorem MAT). Je to však pro komplexní popis anaforických vztahů nedostačující.

Samotná anotace anaforických vztahů v GNOME má tři vrstvy, řazené podle jejich relevantnosti:

1. anotace identické koreference (vztah IDENT, v základním schématu se anotuje pouze tento vztah);
2. anotace mimojazykových deiktických odkazů;
3. anotace asociační anafory.

Asociační anafora se anotuje pouze v případě, když se buď pro daný anaforický výraz nenajde koreferenční antecedent, nebo pokud identický antecedent je ve vzdálenějším kontextu (Poesio 2004a).

Anotace v GNOME má zavedený systém nejednoznačných řešení, které se označují zvlášť.

Při anotaci se dodržují následující pravidla:

- Každý anaforický element musí mít minimálně jeden antecedent, ale ne více, než jeden identický a jeden asociační antecedent.

- Anotuje se vždy vztah s nejbližším předcházejícím antecedentem.
- U anaforického výrazu má být označen každý identický koreferenční vztah, který se najde, ale vztah asociační anafory může být označen jenom jeden.

Tabulka č. 4 stručně charakterizuje některé rysy anotace korpusu GNOME, které mohou být relevantní pro vypracování našeho schématu pro PDT:

	Realizace v GNOME
předběžná anotace „markables“	ANO, všechny NP
velikost korpusu	cca. 1500 vět, cca. 10 tis. anotovaných NP
dodatečné syntaktické, sémantické a diskurzivní informace	*gramatické*: • gramatická funkce, shoda; *sémantické*: • logická struktura (term, kvantifikátor, predikát), • počitatelnost, pomnožnost, • abstraktní-konkrétní, životnost, • genericita, • jedinečnost; *diskurzivní*: • deixe, • kontextová zapojenost (v daném diskurzu nový/dá se vyvodit/v daném diskurzu již známý).
užitečná informace o NP v anotaci GNOME, kterou nemáme k dispozici v PDT 2.0.	• určitost, • pojmenovaná entita (ano/ne)[66], • počitatelnost, • deiktičnost, • generičnost, • logický typ reference (predikativní, distributivní, termový), • funkčnost (situační unika jistého druhu), • ontologický status NP (abstraktní/konkrétní), • struktura referentu.
co se anotuje	• jmenné fráze, • zájmena,

[66] Tato informace se současně připravuje a bude k dispozici do konce r.2009.

	• koordinované struktury (jmenných frází).
co se neanotuje	• slovesa, • klauze, • propozice • elipsy.
anotace predikace a apozice jako koreference	ne
typy anotovaných vztahů (řazeno podle relavantnosti)	• identická koreference (IDENT), • deiktické odkazy, • asociační anafora.
typy asociační anafory	• ELEMENT (množiny), • SUBSET (podmnožina množiny), • POSS (posese a část/celek).
anotační nástroj	XML, MMAX
automatizace anotace	pouze ruční
zaznamenávání ambiguity	ne
mezianotátorská shoda	IDENT – 79.4% označeny oběma anotátory, BRIDGE – 22% označeny oběma anotátory.

Tabulka č. 4: Anotační schéma GNOME

Podobně jako pro korpus GNOME, anotační schéma MATE bylo aplikováno na korpus italských textů VENEX (Poesio a kol. 2008). Podobně jako v GNOME, na textech korpusu VENEX byla provedena bohatá automatická předanotace gramatické, sémantické a diskurzivní informace, výhodou tohoto korpusu je rovněž zpracovaní dialogických textů, které obohacují (ale také velice komplikují) informace o koreferenci. Úplně anotovaný trénovací korpus VENEX se skládá z 30 novinových článků a 6 dialogů. Tato data jsou průběžně doplňována koreferenčními páry, které jsou výsledkem počítačové hry Phrase Detectives (Chamberlain a kol. 2008a, 2008b). Novinkou korpusu VENEX je rovněž zaznamenávání ambiguity v případě, když anotátor vidí více možných koreferenčních interpretací.

Tabulka č. 5 stručně charakterizuje některé rysy anotace korpusu VENEX:

	Realizace v VENEX
jazyk	italština
předběžná anotace markables	ANO, všechny NP, maximálně automatická
velikost korpusu	30 novinových článků + 6 dialogických textů. Průběžně se doplňuje daty z počítačové hry Phrase Detectives (viz výše).
syntaktické, sémantické a diskurzivní informace	= GNOME
co se anotuje	= GNOME
co se neanotuje	= GNOME
anotace predikace a apozice jako koreference	Ne (= GNOME)
typy anotovaných vztahů (řazeno podle relavantnosti)	= GNOME
typy asociační anafory	= GNOME
anotační nástroj	MMAX
automatizace anotace	ano, identifikace kandidátu na anaforický vztah („markables“)
zaznamenávání ambiguity	ano

Tabulka č. 5: Anotační schéma VENEX

II.3.4. Princip anotace koreferenčních vztahů v Müller – Stube (2001)

Autoři článku Müller – Stube (2001) předkládají schéma anotace koreference a asociační anafory (v jejich terminologii *bridging relations*) na materiálu německého Heidelbergského korpusu textů (577 krátkých turistických informací) v rámcích anotačního nástroje MMAX. Daný přístup disponuje přesným a intuitivním rozlišením koreference a asociační anafory. Nutnou podmínkou koreferenčního vztahu (v terminologii Müller – Stube *anafory)* je odkaz obou členů vztahu ke stejné mimojazykové skutečnosti. Tím se do této skupiny dostávají také jmenné fráze s různou hlavou (např. *Katka – dívka* vs. Vieira – Poesio 2000). Uvnitř koreferenčního vztahu se provádí další vnitřní klasifikace na koreferenci přímou (direct anaphora), pronominální (druhý člen páru je zájmeno) a tzv. IS-A (hyponym – hyperonym) relace. Anotace probíhá ve dvou fázích: první fáze je základní, během ní se anotují tzv. relevantní „markables“, tj. pravděpodobné kandidáty na anafory, kam patří především osobní a

ukazovací zájmena a určité deskripce, dále se tyto kandidáty zařazují do množiny elementů s nimi koreferenčních. Rozhodnutí o jediném správném antecedentu daného kandidátu není na dané etapě povinné a probíhá během další fáze, která je už méně přesná a její evaluace se ani nepokládá za nutnou. Kandidátům se připisuje několik atributů, které ovlivňují jejich vstup do anaforického vztahu. Jsou to následující atributy:

- *np form*, rozlišující mezi pojmenovanými entitami, určitými a neurčitými NP, osobními, přivlastňovacími a ukazovacími zájmeny,
- atribut s gramatickou informací typu *person/number/gender,*
- atribut s gramatickou informací typu *subject, object* nebo *other.*

Všechny tyto informace jsou připisovány druhému členu páru.

Asociační anaforou (bridging) se rozumí nekoreferenční sémantický vztah mezi mimojazykovými entitami, které jsou označovány lexikálními jednotkami daného páru. Vyčleňují se následující podtypy asociační anafory: cause – effect (*stavba – dům*), part – whole (*dům – terasa*), a entity – attribute (*manžel – věrnost*). Vyhledávání antecedentu je pro asociační anaforu obligatorní již v první fázi anotace, ale může se volně vybírat mezi všemi elementy koreferenčního řetězce antecedentu, protože je to v chápání autorů vztah mezi objekty reality, nikoliv mezi lexikálními jednotkami.

Přistup Müller – Stube je stručně představen v tabulce č. 6.

	Relizace v Muller – Stube
jazyk	němčina
předběžná anotace „markables“	ano
velikost korpusu	neuvedeno
evaluace	neuvedeno
anotace asociační anafory	ano
typy asociační anafory	• part – whole (*dům – terasa*), • cause – effect (*stavba – dům*), • entity – attribute (*manžel – věrnost*).
anotace identické koreference	ano
klasifikace identicky koreferenčních vztahů	ne
anotační nástroj	MMAX

Tabulka č. 6: Anotační schéma Müller – Stube

II.3.5. Anotace koreference a asociační anafory v Chiarcos – Krasavina (2005)

Koreference a asociační anafora pro němčinu a angličtinu se zpracovává v rámci projektu PoCoS (Postdam Commentary Corpus). Anotační schéma je představeno v manuálu anotace RST Discourse Treebank (Carlson a kol. 2003) a korpusu německých komentářů Postdam Commentary Corpus (Stede 2004). Anotace je rozdělena na dvě části – základní anotace koreference, teoreticky jazykově nezávislá, s omezeným počtem příznaků, lehce měnitelná a adaptovatelná k novým cílům a zavedení/změně příznaků, vhodná pro budoucí automatické zpracovaní koreference a jiné experimenty z oblasti počítačové lingvistiky. Rozšířená anotace koreference je zpracována pro konkrétní jazyk (angličtinu a němčinu), má vyšší ambiguitu a méně přesnou sémantiku typů, je méně vhodná pro automatické zpracování, ale obsahuje z lingvistického hlediska více informace, může být tedy vhodnější pro lingvistické výzkumy. Předmět anotace anaforických vztahů („markables“) se dělí na primární a sekundární. V základní verzi anotace primární jsou osobní zájmena (vyjádřená v textu), určité a posesivní deskripce, pojmenované entity a názvy, pronominální adverbia; sekundární jsou hlavně neurčité deskripce, které se označí, pokud vystupují v textu jako antecedent v anaforickém vztahu. V rozšířené verzi k základním jednotkám přibývají tázací, reflexivní a nulová zájmena, k sekundárním – propozice. Anotují se koreferenční anaforické vztahy a kataforické vztahy v

rámci jedné věty. V rozšířené verzi se tyto vztahy klasifikují na modifikace (vyjádření jinými slovy, dodávání nové informace), synonymie, opakování stejné nebo skoro stejné NP (*der Kanzler ... der Kanzler ... Der Bundeskanzler*) a pronominalizaci. Rovněž v rozšířené verzi se anotuje asociační anafora, jejíž klasifikace se opírá na systém typů v Gardentovi (2003). Při anotaci se postuluje řada konvenčních preferencí, jak anotovat některé typy případů, pokud existuje několik možností. Technický nástroj pro anotaci je MMAX. Anotace se provádí na textu (nikoliv na stromech) a teoreticky vychází ze složkového principu, i když nemají syntaktickou strukturu v podobě stromu.

Anotační proces v PoCoS zahrnuje řešení velkého počtu atributů. Je to pravděpodobně časově mnohem náročnější anotace, než probíhající anotace koreference na PDT, kterou popisujeme v dané práci. Anotátor označuje nejenom typ koreference, ale také atributy „direct speech", typ fráze (NP, PP, jiná), typ NP (named entity, určitá NP, neurčitá NP, osobní zájmeno apod.) a atribut „typ ambiguity", který má až 7 významů. V atributech rozšířeného schématu zpracování koreference anotační schéma PoCoS obsahuje informaci o sémantické třídě (abstraktní, osoba, materiální objekt, událost apod.) Něco je derivováno z WordNetu. Srovnáme-li to s PDT, je část těchto informací zahrnuta v tektogramatických atributech PDT 2.0. Pozitivní na anotaci v PoCoS je však to, že každý atribut má význam „other", kam anotátor může umístit nejasné příklady. Na druhé straně je to také poměrně nebezpečné – není vyloučeno, že velký počet případů bude "vyhozen" do typu *other* a klasifikace tím výrazně ztrátí na významu.

Zvlášť je zaveden systém typů pro nejasné případy. Např. v případě nejednoznačného výběru antecedentu, anotátor označí daný příklad atributem *ambig_ante*, a na tento uzel by neměla vést žádná koreferenční šipka (aby se předešlo nejasnostem). Takové řešení se zdá být velice přínosné, protože pak „čisté" případy jsou vyznačeny a všechny evaluace se mohou provádět pouze na nich.

Nehledě na to, že rozšířená verze projektu PoCoS ještě nebyla implementována na velkém textovém korpusu (a ani už pravděpodobně nebude), koncepce anotace je velice zajímavá z různých hledisek. V tabulce č. 7 upozorňujeme na několik základních rysů anotačního schématu PoCoS:

	Realizace v PoCoS
jazyk	angličtina a němčina
předběžná anotace „markables“	ano
co se anotuje	• jmenné fráze, • předložkové fráze, • koreference na propozice (má se anotovat v rozšířené verzi), • adverbia, zadaná seznamem (*there, then* apod).
co se neanotuje	• adjektiva (v žádné syntaktické funkci)
řešení koordinovaných struktur	NP/PP v koordinačních konstrukcích jsou anotovány dvakrát – jako celek a zvlášť elementy.
koreference vs. anafora	Anotují se anaforické vztahy, nikoliv koreference.
gramatická koreference	Gramatická koreference, která se v PDT řeší zvlášť (Kučová a kol., 2003), je v PoCoS rozdělena na základní a rozšířenou, přičemž reflexivní zájmena se řeší až v rozšířené verzi, tj. nejsou myšlena pro budoucí automatické zpracování.
řešení ambiguity	ano, v atributu ambig
identifikace vs. predikace v konstrukcích jmenných se sponou	Rozlišení se neprovádí, ale upozorňuje se na něj v manuálu.
k dodržování koreferenčních řetězců	V anotaci koreference v rámcích základního schématu se dodržuje koreferenční řetězec. V rozšířené anotaci však koreferenční jednotky nemohou sloužit jako antecedent pro anaforické vztahy, aby se následně mohlo propojit základní a rozšířenou anotaci. Tím se však „zbavuje“ koreferenčních řetězců.
sémantický koncept antecedentu	V rozšířené anotaci koreference jako antecedent slouží význam celého souborů předchozích antecedentů, tedy informace se množí a doplňují se. Pro určení typu vztahu se dívá na celý předchozí kontext, nikoliv jenom na poslední antecedent.
anotace asociační anafory	ANO (má se anotovat v rozšířené verzi), přičemž asociační anafora se neanotuje, pokud NP jsou významově spojené nikoliv kontextem, ale obecnou znalostí světa. To je poměrně vágní definice, protože

	jako kontext je v Chiarcos – Krasavina (2005) chápána i slovníková informace. Spíše se dá říci, že se neanotují jenom okazionální implicitní vztahy. • Asociační anafora se anotuje jenom u slov s plnohodnotnou lexikální sémantikou – žádná zájmena a elipsy; • Asociační anafora se neanotuje pokud jeden z členů je součástí přímé řeči, zatímco druhý není.
identita vs. asociační anafora	Asociační anafora se neanotuje, pokud jakékoli části daných NP jsou spojené identickou koreferencí. Tedy tento vztah se chápe jako poměrně sekundární – anotuje se pouze v případě, když to zůstává úplně nepropojené.
anotační nástroj	MMAX

Tabulka č. 7: Anotační schéma PoCoS

II.3.6. Projekt anotace koreference AnCora-CO pro španělštinu

Podrobný algoritmus anotace koreferenčních vztahů je představen v prácích M. Recasensové (2008, 2010) na materiálu korpusu španělských textů AnCora-CO. Korpus se skládá z textů novinových článků a obsahuje cca 500 tisíc slov oanotovaných na morfologické, syntaktické a sémantické rovině. Koreferenční vztahy se anotují u plných jmenných frází, zájmen a aktuálních elips. Vedle referenčních jmenných frází jsou do koreferenční anotace zahrnuty atributivní a predikační NP, které se však anotují zvlášť, ne společně s referenčním použitím. Také se rozdílně anotuje koreference výrazů se specifickou a generickou referencí. Zvláštní pozornost je věnována koreferenčním vztahům mezi pojmenovanými entitami, které jsou vyčleněny a klasifikovány již na sémantické rovině původní anotace AnCora. Speciální vztah je zaveden pro tzv. *discourse deixis*, který je u Recasensové chápán jako odkazování k situaci (větě nebo klauzi) nebo většímu úseku textu. Asociační anafoře v tomto projektu je věnován článek Recasensová a kol. (2007), ve kterém se autoři pokouší aplikovat systém meta-schématu MATE na španělský korpus, avšak na velkém korpusu anotace provedena zatím nebyla.

Anotační postupy M. Recasensové shrnuje tabulka č. 8:

	Realizace u Recasensové (2008)
jazyk	španělština
předběžná anotace „markables“	ano, částečně automatická
co se anotuje	• jmenné fráze, • předložkové fráze, • zájmena, • aktuální elipsy, • atributivní a predikační NP, • „discourse deixis“.
co se neanotuje	• adjektiva
asociační anafora	plánuje se, ale ještě nebyla aplikována na velkém korpusu
typy asociační anafory	= MATE (GNOME)
anotační nástroj	PALinkA

Tabulka č. 8: Anotační schéma AnCora-CO

II.3.7. Jiná anotační schémata

Anotace pouze pronominální identické koreference je představena např. v **Tutinovi** (2000) na velkém (milion slov) korpusu francouzských textů. Z důvodů velké subjektivity výběru anotovaných jednotek se autoři omezili na anotaci anaforických vztahů u osobních, přivlastňovacích a ukazovacích zájmen, anaforických adverbií, elidovaných členů a některých specifických konstrukcí. Provádí se rovněž referenční klasifikace, tj. rozdělení na případy koreference, vztahy množiny – podmnožiny, rozděleného antecedentu apod. Evaluace se provádí pouze zkušebně, bez uvedení procentuálního vyjádření.

Minoz (2000) na materiále španělštiny zkouší algoritmus vyhledávání asociační anafory a koreference pomocí tezauru sémantických vztahu WordNet. Termíny asociační anafora a koreference jsou definovány poměrně vágně, asociační anafora zahrnuje všechny různé nominace, i koreferenční, ve výkladu samém je pak vše označeno jako koreference. O anotaci se v článku nepíše, ale pravděpodobně byla provedena pro výpočet evaluace. Ačkoliv nejsou použita manuální pravidla, výsledky jsou velice slibné: 60.9% pro precision a 78% pro recall.[67]

Jako příklad automatického vyhledávání asociační anafory uvedeme práci (**Vieira – Teufel 1997**), kde je představen pokus o automatické zpracování na materiálu 20 článků z Wall Street

[67] K pojmům *precision* a *recall* viz např. Makhoul a kol. 1999.

Journal. Autoři projektu mají podstatně jiné rozdělení koreference na vztah identity a asociační anafory, než používáme my pro anotaci PDT. Jako asociační anafora se rozumí rovněž vztah synonymie a hyponymie – hyperonymie u koreferenčních jmen.

Autoři oprávněně podotýkají, že pro automatické zpracovaní asociační anafory je třeba zapracovat dostatečný slovník. Pro tyto potřeby byla použita veřejně přístupná lexikální databáze WordNet (Müller 1998). S použitím této databáze byly provedeny experimenty identifikace asociační anafory. Ukázalo se, že slovník pomáhá odhalit pouze 19 procent vztahů, označených ručně anotátory. Budou to především vztahy synonymické, hyponymické (hyperonymické) a vztahy typu část – celek. Neodhalují se především vztahy mezi pojmenovanými entitami a určitými deskripcemi (jako např. *Mrs. Park – the housewife* nebo *Pinkerton's Inc – the company*), anafora na větu nebo slovesnou frázi (*Kadane Oil Co. is currently drilling two wells... – The activity...*), složené deskripce, pro které je důležitý nejenom řídicí uzel, ale také všechna jeho určení (*stock market crash – the markets*, nebo *discount packages – the discounts*), sémantické vztahy návaznosti, důsledky, množiny – podmnožiny a některé jiné, které výrazně přispívají ke koherenci textu. Kromě toho nebyly odhaleny některé vztahy typu synonymie, hyponymie a části – celku, které by teoreticky mohly být rozpoznány jako např. synonymické: *new album – the record, three bills – the legislation*; hyperonymické *rice – the plant, the television show – the program, m*eronymmické *plants – the pollen, the house – the chimney* aj.

Jiný problém, který se vyskytl při automatickém vyhledávání vztahů asociační anafory, je velký počet nalezených vztahů, které nejsou správné nebo nutné. Tak se např. najde vztah *Mrs. Housman – 50 years old.*

II.4. Celkové zhodnocení teorií a korpusů

Představené v II.1.– II.3. práce popisují state of the art v oblasti teorie reference a možnosti její aplikace především na manuální zpracování koreference na velkých textových korpusech, která má sloužit jako tréninkový materiál při pokusech o automatické zpracování daného jevu. V této kapitole jsme nepojednali o experimentech automatického zpracování anafory a koreference (anaphora and coreference resolution) samotných. Ačkoliv je tato oblast velice důležitá, je mimo rámce našeho anotačního projektu.

Popsaná teoretická báze tvoří základ následující analýzy. Jako východisko pro rozlišení typu reference výrazů ke skutečnosti jsme vybrali především klasifikaci Padučevové (II.2.) upřesněnou T. Bergerem a I. Mendozovou (viz podrobnější rozbor v III.4.2.1.). České teorie

Palka a Hlavsy se s nimi v podstatě shodují, mají však složitější terminologický systém a příliš teoretické zaměření, které je těžko aplikovatelné na velké textové korpusy se skutečnými texty. Velký význam pro nás měla klasifikace anaforických vztahů u Daneše, kterou jsme využili při vypracování typologie asociační anafory.

Výzkumné projekty, jejichž součástí je rovněž zpracování koreference a asociační anafory se různí v přístupech k anotaci, cílech, velikosti hotových anotovaných korpusů a (pokud byla změřena) v míře mezianotátorské shody. Rozpracovaná anotační schémata v II.3.1.–II.3.7. jsou výběrem takových projektů, které nám posloužili pro vypracování vlastního schématu. MUC (II.3.1.) a ACE (II.3.2.) jsou neznámějšími a nejpoužívanějšími schématy pro trénink automatických metod, jsou však velice omezena co do typů a počtu vztahů, které zaznamenávají. Chápeme je především jako měřítko toho maxima, které pro daný moment může být zpracované automaticky, a snažíme se od něj nevzdalovat. Anotační schéma v PoCoS (II.3.5.) je naopak velice podrobné a výsledky jeho aplikace se zdají být zajímavé jak z teoreticky-lingvistického tak i z počítačově-lingvistického hlediska. Jeho důslednost a podrobnost při zpracování konkrétních jazykových příkladů byly jedním z hlavních orientačních bodů při vypracování našeho schématu pro PDT. Bohužel však anotační schéma v PoCoS nebylo (z provozních důvodů, ne kvůli jeho nadměrné složitosti) použito pro anotaci velkého korpusu textů. Pravděpodobně největší vliv na typologii vztahů mělo meta-schéma MATE (II.3.3.) a jeho modifikovaná aplikace na španělštinu od M. Recasensové (II.3.6.). Poměrně podrobná klasifikace typů koreference a asociační anafory je spojená se snahou o co největší formální přístup a kritickou analýzu mezianotátorských neshod a jejich řešení. Velký počet problémů, na které narážejí vědci pracující v těchto projektech jsou velice podobné problémům, které vznikají i během naší anotace koreference a asociační anafory na PDT. Ne všechny tyto problémy jsou v MATE a Recasensové (2007) vyřešeny, avšak jejich konstatace v několika různých projektech svědčí o tom, že tyto problémy skutečně existují v jazyce a ve formálním přístupu k němu, tedy společná práce z různých stran může přispět k jejich vyřešení v zohlednitelné budoucnosti.

III. Schéma anotace rozšířené koreference na PDT

V oddíle II jsme vymezili teoretické aspekty analýzy anaforických vztahů a koreference, ke zkoumání kterých chceme přispět anotací velkého textového korpusu. V této části práce se pokusíme o zpracování praktického schématu anotace rozšířené koreference a asociační anafory na tektogramatické rovině v PDT.

Ve stávající anotaci koreference a asociační anafory na PDT rozlišujeme následující tři typy vztahů:

- gramatická koreference (v příkladech zkracujeme na *coref_gram*) – koreferenční vztahy uvnitř jedné věty, které jsou přesně determinovány gramatickými pravidly daného jazyka, viz III.3.,
- textová koreference (v příkladech zkracujeme na *coref_text*) – ostatní případy koreference, viz III.4.,
- asociační (bridging) anafora (v příkladech zkracujeme jako *bridging*) – sémantický vztah bez koreference, viz III.5..

Gramatická koreference byla již oanotována na celém korpusu PDT,[68] textová koreference je anotována v původní verzi částečně (III.4.1.), vztahy asociační anafory zavádíme nově.

[68] Viz Kučová a kol. (2003)

III.1. Teoretické zásady anotace

V dané kapitole uvádíme některé teoretické principy, konvence a zásady, které dodržujeme při anotaci koreferenčních vztahů a asociační anafory na PDT.

III.1.1. Princip důslednosti

Princip důslednosti nebo maximální nezávislosti na subjektivním názoru anotátora je velice důležitý pro dodržení jednotnosti anotace (tzv. mezianotátorské shody). Pravidla mají být formulována s maximální přesností a detailností.

Tento princip se relativně snadno dodržuje v případě gramatické koreference, kde pravidla anotace jsou determinována gramatickými pravidly jazyka a jsou většinou nezávislá na subjektivním názoru anotátora. Lingvistická podstata anotovaných jevů však vede k tomu, že se dodržování daného principu zeslabuje směrem od gramatické koreference k asociační anafoře. V případě gramatické koreference je možnost rovnocenného výběru mezi více antecedenty pro jeden anaforický člen prakticky vyloučena – gramatická pravidla jazyka (z definice gramatické koreference) předurčí pouze správný antecedent. Avšak pronominální textová koreference, která již byla oanotována na celém PDT, případy více interpretací nevylučuje; je jich však ještě poměrně omezený počet a dá se stanovit relativně přesná pravidla výběru toho či onoho řešení.[69] Rozšířená textová koreference a asociační anafora, kterým se věnujeme v této práci, jsou v tomto ohledu nejvíce problematické. Porušení důslednosti mohou být trojího druhu:[70]

- nedostatečně přesně formulovaná pravidla anotace – minimalizuje se stanovením přesnějších pravidel pomocí této práce;
- možnost více interpretací, které je anotátor schopen navrhnout – řeší se určováním preferencí výběru nebo možností označování těchto případů jako ambiguitních (srov. např. v Chiarchos – Krasavina 2007);
- textová ambiguita, kterou si anotátor většinou neuvědomuje. Je přesvědčen, že textu rozumí správně, a že jeho interpretace je jediná možná. Při srovnání několika anotací však vychází najevo, že i jiné interpretace jsou stejně oprávněné. V tomto případě přesná formální pravidla již nestačí pro dodržení důslednosti anotace. Takový typ ambiguity nejsme schopni zatím vyřešit.

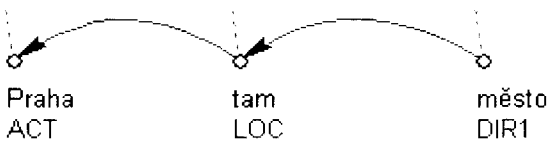
[69] Ibid.

[70] Pro příklady a rozbor problematických neshod těchto typů viz IV.4.

Přestože ne všechny situace se zdají v danou chvíli řešitelné, je snaha o dodržení důslednosti anotace naším zásadním úkolem a celá tato práce by k tomu měla přispět.

III.1.2. Princip dodržování (maximálního) koreferenčního řetězce

Při anotaci všech typů koreferenčních vztahů se dodržuje zásada udržovat jednoduchou linearitu koreferenčního řetězce. V případě více interpretací odkazujeme vždy na nejbližší předcházející koreferenční uzel. Tedy pokud A, B a C jsou po sobě následující jmenné fráze, přičemž C je koreferenční s A a B, koreferenční šipka vede od C k B a dále od B k A. Srov. obrázek č. 2, zobrazující posloupnost (*Praha – tam – město*)



Obrázek č. 2: Dodržování koreferenčního řetězce pro textovou koreferenci

Praha - tam - město

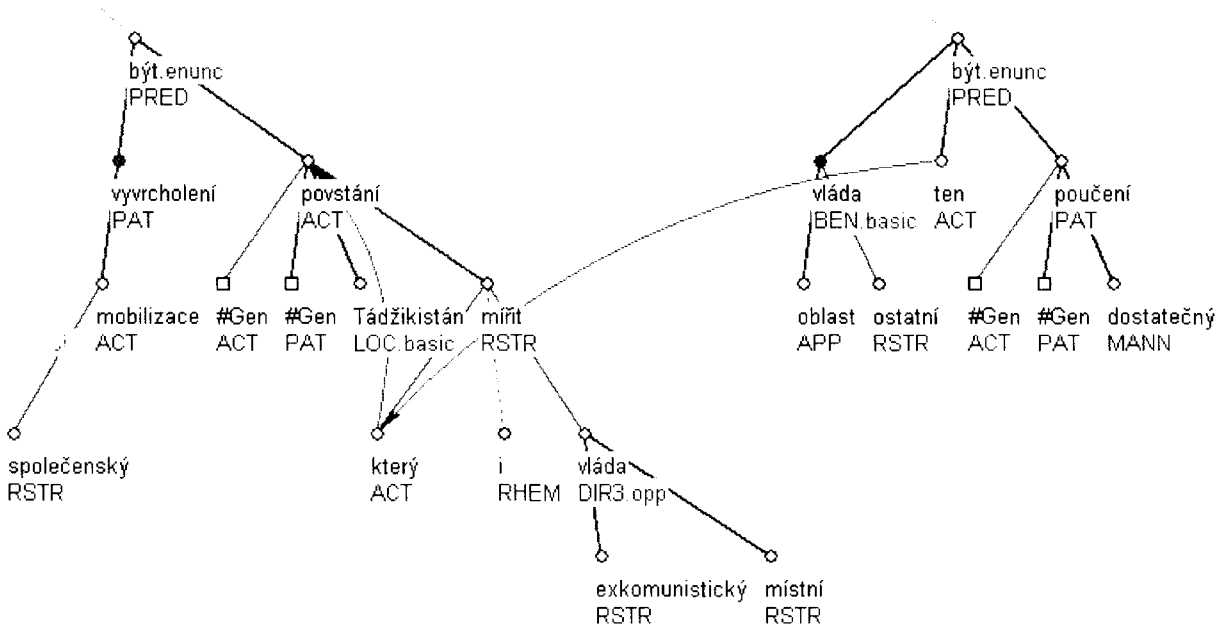
Dodržování koreferenčního řetězce je kontrolováno automaticky. Pokud anotátor nakreslí šipku na uzel, na který už vede šipka identické koreference, jeho šipka se automaticky překreslí na poslední uzel daného řetězce.

Kromě dodržování koreferenčního řetězce mezi šipkami identické textové koreference se udržuje také nepřetržitost řetězce mezi šipkami gramatické a textové koreference.[71] Tedy pokud A, B a C jsou po sobě následující jmenné fráze, přičemž B je propojeno s A gramatickou koreferencí a C je textově koreferenční s A, šipka textové koreference vede od C k B. Srov. obrázek č. 3, kde od ukazovacího zájmena *to* odkazujícího na *povstání* vede šipka textové koreference na vztažné zájmeno *který* ve větě (1)a.

(1) *a. Vyvrcholením společenské mobilizace bylo povstání v Tádžikistánu, které* {coref_gram na „povstání"} *mířilo i proti místní exkomunistické vládě.*

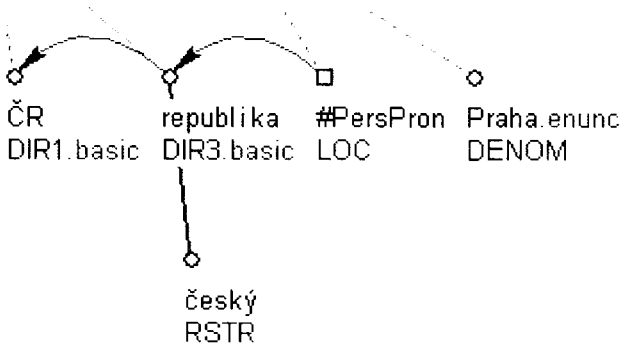
[71] V původní anotaci koreference na PDT se snažilo o udržování nepřetržitého řetězce mezi šipkami gramatické a textové koreference, tento princip však nebyl dodržen pravidelně. Např. v souboru train-2 ze 302 případů, koreferenční řetězec udržují pouze 242 případů (80%). Při anotaci rozšířené koreference jsme tuto nepravidelnost v anotaci původní pronominální koreference automaticky opravili tak, aby se koreferenční řetězec dodržovat ve všech případech.

b. Pro ostatní vlády oblasti <u>*to*</u> {coref_text na „který“ v *a.*} *bylo dostatečným poučením.*



Obrázek č. 3: Dodržování koreferenčního řetězce mezi gramatickou a textovou koreferencí

V případě asociační anafory šipka vždy vede na nejbližší předcházející uzel koreferenčního řetězce antecedentu. Anaforické páry s lexikálně vyjádřenými uzly je možné podle potřeby vyhledat automaticky. Srov. na obrázku č. 4 zobrazujícím posloupnost (*ČR – Česká Republika – s ní – Praha*) vztah část – celek mezi Českou Republikou a Prahou bude označen mezi „Praha“ a „s ní“. Tato zásada je dodržována automaticky, čili ať už anotátor vede šipku na jakýkoli uzel koreferenčního řetězce, skutečná šipka vždy povede na poslední uzel daného řetězce (viz IV.1.2.2.).



Obrázek č. 4: Dodržování koreferenčního řetězce: asociační anafora

ČR - Česká Republika - s ní - Praha

Srov. také (2)a–b, kde anotátor s největší pravděpodobností spojí asociační anaforou *kluci a děvčata* a *několik*, automaticky však skutečná šipka povede od *několik* na *který*.

(2) *a. Na toto telefonní číslo však mohou samozřejmě zavolat všichni kluci a děvčata, kteří* _{coref_gram na „kluci a děvčata"} *se ocitnou ve svízelné situaci.*
b. Ptali jsme se několika {bridging na „který" v *a.*}, *jestli by takového kamaráda po telefonu považovali za dobrou věc.*

III.1.3. Princip maximální velikosti koreferující jednotky

Za člen koreferenčního páru se považuje celý podstrom koreferujícího výrazu, tj. koreferující člen vždy zahrnuje determinátory (ukazovací zájmena apod.), modifikátory (uzly s RSTR apod.), předložky a všechny závislé členy. Takové řešení je částečně podmíněno strukturou tektogramatického stromu, kde nelze jednoduchým způsobem vyloučit z koreferenční entity některé závislé uzly a zůstat přitom na tektogramatické rovině. Avšak tento princip dodržujeme i v jiných případech (např. neodkazujeme na uzly s funktorem ID (III.4.2.4.1.), naopak při výběru mezi kontejnerem a jeho závislým členem odkazujeme na kontejner (III.4.2.3.3.), v apozičních a koordinačních konstrukcích odkazujeme na spojku (III.4.2.4.1., III.4.2.4.2.) atd.) Princip má ten nedostatek, že se někdy do koreferujícího podstromu dostávají jednotky, které se vztahu koreference nezúčastní (viz III.4.2.3.5.1.). To však nelze vyřešit bez podstatné změny technického zázemí anotace, proto ve stávající anotaci ponecháváme tento problém stranou. Kromě toho vyloučení některých závislých uzlů z

koreferenčního vztahu způsobí větší rozdíly mezi anotátory, což si v dané etapě nemůžeme dovolit (k mezianotátorské shodě viz IV.3.).

Princip maximální velikosti koreferující jednotky však nevylučuje, že část koreferujícího podstromu může koreferovat samostatně v rámci jiného koreferenčního vztahu.

III.1.4. Princip kooperace se syntaktickou strukturou tektogramatické roviny

Syntaktická struktura anotovaných stromů PDT umožňuje zjištění některých koreferenčních vztahů automaticky. Platí tedy princip, že pokud je koreferenční vztah zřejmý ze struktury stromu, tak ho neanotujeme.

Z těchto důvodů neanotujeme:

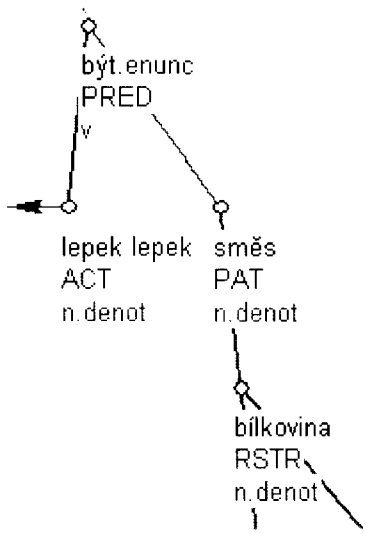
1. Vztah **mezi jednotlivými členy apozice.** Z podstaty apozice vyplývá, že jednotlivá pojmenování jsou koreferenční. Apoziční vztah je zachycen v tektogramatickém stromě funktorem APPS a syntaktickou strukturou. Vztah mezi členy apozice neoznačujeme ani pro identickou koreferenci (*Božena Němcová* .APPS *autorka Babičky*; *slečna Sollárová, jinak.* APPS *slovenská malířka*; *ODS (Občanská demokratická strana)*), ani pro asociační anaforu (srov. (3)–(4)).

(3) *Tomu odpovídala cílová místa – Kypr, Kréta, Malta* {žádný koreferenční vztah}.

(4) *Přijeli do měst, jako Praha, Brno a Ostrava* {žádný koreferenční vztah}.

2. Vztah **mezi subjektem a predikátovou částí výpovědi** ve větách s predikátem jmenným se sponou (*Petr je lékař – Petr ... lékař*). Predikace je usouvztažnění jednoho slova k druhému, skutečnosti ke skutečnosti, takže anotace koreference mezi subjektovou a predikátovou částí by byla zbytečná. Tento vztah je reprezentován v syntaktické struktuře stromu. Případnou koreferenční šipku vedeme na subjekt/od subjektu. Srov. segment tektogramatické reprezentace věty (5) na obrázku č. 5, kde neanotujeme koreferenční vztah mezi uzly *lepek* a *směs*.

(5) *Lepek* {coref_text na „lepek" v předchozím kontextu} *je směs* {žádný koreferenční vztah} *ve vodě rozpustných bílkovin z povrchní části obilných zrn pšenice, žita, ječmene i ovsa.*



Obrázek č. 5: Koreference mezi subjektem a predikátovou částí výpovědi

Koreferenční vztah mezi subjektem a predikátovou částí výpovědi neoznačujeme ani v případě konstrukcí, kde subjektem je ukazovací zájmeno *to*. Koreferenční vztah s předchozími/následujícími uzly se připojí na zájmeno. Srov.(6)a–b:

(6) *a. My tady máme dost problematických dětí a ty kdyby se spolčily s Němci, to by nedělalo dobrotu.*
b. „Slyšeli jsme, že to {coref_text na „Němci"} *jsou děti* {žádný koreferenční vztah}, *které místo výkonu trestu mají být tady v Košťanech..." říká obyvatelka obce a signatářka petice.*

V identifikačních větách, kdy oba členy predikačního vztahu mají vlastní referenci, můžeme výjimečně podrobit anotaci obě části predikačního vztahu, ale přesto neanotujeme koreferenci mezi subjektem a jmennou částí predikátu – jejich koreference je dána syntaktickou strukturou stromu a může být podle potřeby doplněna automaticky. Srov. (7)a–c:

(7) *a. Prvotní apoštolská církev byla chudá. Přesto i ona měla jakousi finanční organizaci, dokonce svého pokladníka.*
b. Problémem je, že tímto prokazatelně prvním křesťanským ekonomem {coref_text na "pokladník"} *byl Jidáš Iškariotský* {žádný koreferenční vztah}.
c. Neblahé stigma Ježíšova zrádce {coref_text na "Jidáš" v *b.*} *jako by se nad*

církevním majetkem vznášelo dodnes.

Rozdělení syntaktických konstrukcí s predikátem jmenným se sponou na predikační a identifikační však není úplně bezproblémové (viz k tomu teoretický výklad v II.2.1.). Pokus o rozdělení identifikace a predikace v anotaci koreferenčních vztahů byl proveden v projektu MATE (II.3.3.) V původní verzi se neoznačoval koreferenční vztah mezi subjektem a jmennou frází ve jmenné části predikátu ve větách typu *John is a policeman*, zatímco se musel označovat v identifikačních větách typu *The planet on the left is Venus*, kde jmenná fráze v predikátu má specifickou referenci. Ukázalo se však, že rozlišování mezi predikující a referující NP v pozici predikátu není intuitivní, zvlášť v jazycích jako je italština (rovněž řešená daným projektem), kde se subjekty v takových konstrukcích často používají predikativně.[72]

3. Neanotujeme koreferenci u atomických uzlů,[73] uzlů reprezentujících cizojazyčné výrazy a závislé části frazeologických spojení. Žádný vztah rovněž nevede ke kořenu tektogramatického stromu.

4. Neanotujeme vztah u komplexních uzlů s některými funktory:

   a) Podobně jako při anotaci původní pronominální koreference, v rozšířené anotaci textové koreference a asociační anafory v případech, kdy antecedent je substantivní skupina s tzv. nominativem jmenovacím (výrazem s funktorem `ID`) nevede koreferenční vztah většinou k doplnění v pozici nominativu jmenovacího, ale k jeho řídícímu substantivu; za koreferovaný člen považujeme proto řídící uzel uzlu s funktorem `ID`.

   b) Asociační anafora se neanotuje mezi uzly s funktory `APP`, `MAT`, `AUTH` nebo PAT) a jejich řídícím uzlem (viz III.5.2.3.).

III.1.5. Preference koreference před asociační anaforou

Pokud můžeme vybírat mezi identickou koreferencí a asociační anaforou, vybíráme identickou koreferenci, i v případě kdy identický antecedent je v předcházejícím kontextu vzdálenější. Srov. v (8)a–c jmenná fráze „žvýkací guma" v (8)c odkazuje k „žvýkačka" ve větě (8)a. Ačkoliv mezi (8)a a (8)c je věta (8)b se stejnou jmennou frází „žvýkací guma", ta není

[72] Příklady a vysvětlení viz v Poesio (2004a).

[73] K vysvětlení termínů viz Mikulová a kol. 2005, s. 14n.

koreferenční se „žvýkací gumou" v (8)c, protože v (8)b „žvýkací guma" má specifickou referenci, kdežto v (8)a a (8)c generickou:

(8) *a. Vybrané kapitoly z dějin žvýkačky.*
b. Jeho milovaný kousek žvýkací gumy {bridging na „žvýkačka" v *a.*}, *který si tak pečlivě odložil na spodek desky stolu , se stal kořistí nepřítele a asi jej čeká potupný konec v odpadním koši.*
c. Pro historii žvýkací gumy {coref_text na „žvýkačka" v *a.*}, *jak ji známe dnes, se však musíme přenést na jiný kontinent.*

V případech, kdy jmenná fráze odkazuje k antecedentu vztahem asociační anafory, přičemž je to důležité pro koherenci textu (viz III.1.7.), a zároveň má (většinou ve vzdálenějším kontextu) textově koreferenční uzel, označujeme oba vztahy – asociační anaforu i textovou koreferenci. Srov.:

(9) *a. Jistotu v tomto směru dávají nejnovější kroky vlády SR, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience.*
b. Byť má na tento krok {coref_text na „zavést", bridging na „nejnovější kroky"} *určité právo (jako člen GATT), v daném okamžiku však vyznívá jako tvrdé politické rozhodnutí vlády, která se snaží velice rezolutními administrativními kroky zredukovat mnohamilionové pasívum v obchodní výměně s ČR.*

III.1.6. Princip rozhodujícího koreferenčního vztahu

Koreferenční vztah není totéž co anaforický. Koreference je reference dvou výrazů na tentýž objekt reálného světa, zatímco anafora je odkaz na to, co bylo již v předchozím kontextu zmíněno. Z existence koreference mezi dvěma členy páru automaticky neplyne, že mezi nimi existuje anaforický vztah, i když tomu tak často bývá. Ve stávající anotaci bereme za základ vztah koreferenční, nikoliv anaforický, tedy zaznamenáváme koreferenci i v případě, že anaforický vztah chybí. Anaforické vztahy bez koreference anotujeme jako asociační anaforu (viz III.5.1.5.1.).

S neanoforickou koreferencí se nejčastěji potkáváme v případě, že v textu nejsou žádné další prostředky koheze a je splněna alespoň jedna z následujících podmínek:

1. Oba členy koreferenčního páru se nachází v rématu. Srov. koreferenční NP *dobrovolníci* a *volontéři* ve dvou po sobě následujících větách (10)a a (10)b:

(10) *a. Příjemně ji překvapilo , že se přihlásilo tolik dobrovolníků, kteří chtějí pomáhat druhým lidem.*
b. Nyní má linka třicet osm tzv. volontérů {coref_text na „dobrovolník"}, *kteří budou naslouchat volajícím.*

2. Členy koreferenčního páru jsou ve stejném textu, ale v různých diskursivních jednotkách. Srov. NP *Košťany* a *obec Košťany na Teplicku* v (11) a–b:

(11) *a. Před několika týdny zaplnily stránky regionálních deníků i celorepublikových časopisů články, jejichž titulky „Jugend prý nabízí dětem alkohol a svádí patnáctileté dívky" nebo „Většina obyvatel by zřejmě mezi sebe problémové děti, které k nám vozí na převýchovu německá církev, nepřijala" a „Němečtí problémoví odešli z Košťan" naznačovaly odhalení skandálu.*
b. V obci Košťany na Teplicku {coref_text, na „Košťany" v *a.*} *ještě chlapci ani nebyli, ale místní již dali dohromady petici: „My rodiče dětí základní školy Košťany protestujeme proti umístění ubytovny pro potrestané německé chlapce."*

3. Členy koreferenčního páru jsou od sebe ve větší textové vzdálenosti:

(12) *a. Podle těchto zpráv nějaká firma na naše území umísťuje německou delikventní mládež, která zde páchá kriminální činy a ohrožuje starousedlíky.*
[... 15 vět ...]
b. V Košťanech totiž zakoupila dům firma Struktura {coref_text na „firma" v *a.*}, *která se u nás rozmísťováním německých chlapců zabývá.*

Ve všech uvedených příkladech (10)–(12) koreferenci vždy označíme, protože odpovídající jmenné fráze jednoznačně odkazují na stejný mimojazykový objekt, avšak je jasné, že nejde o anaforický odkaz, neboť *volontérů* v (10)b ani *obci Košťany na Teplicku* v (11)b a *firma Struktura* v (12)b neobsahují odkazování k odpovídajícím koreferenčním NP v (10)a–(12)a.

Dodržování zásady orientace na koreferenci nikoliv na anaforu není zdaleka samozřejmostí. Většina anotačních schémat, které zpracovávají koreferenční vztahy na angličtině, berou za základ právě anaforu, nikoliv koreferenci. Orientace na anaforu má také jisté pozitivní stránky

– vždyť právě především anaforické odkazování zaručuje koherenci textu a koherence textu je právě to, co chceme ve výsledku zachytit v anotovaném korpusu. Avšak angličtina má unifikovaný prostředek vyjadřování určenosti (také kontextové zapojenosti, známosti) jmenných frází v textu – určitý člen – a také formální gramatickou kategorii určenosti, která činí toto vyjadřování obligatorním. Tato skutečnost umožňuje pro angličtinu (i pro jiné jazyky s gramatickou kategorií určeností) bez větších formálních potíží vybírat definitivní jmenné fráze v textu a hledat jejich antecedenty. Najdou se sice definitivní jmenné fráze, které v předchozím kontextu žádný antecedent nemají, je jich však menšina a podléhají dostatečně formální klasifikaci.[74] Čeština jako jazyk bez gramatické kategorie členu má také jisté prostředky pro vyjadřování určenosti. Jsou to např. pronominalizace, elipsa, ukazovací zájmena, aktuální členění, intonace, slovosled, slovesný vid apod. V případě pronominalizace, elipsy a některých případů opakování s ukazovacími zájmeny je možné anaforický vztah s velkou pravděpodobností předpokládat. Z toho důvodu byl také zachycen v původní anotaci pronominální koreference. Jakmile se však dostáváme na úroveň lexikálně vyjádřených jmenných frází, situace se výrazně komplikuje. Nemáme k dispozici přesný formální mechanismus vyčleňování definitivních NP v textu. Tomuto tématu se sice věnuje řada teoretických lingvistických studií,[75] jsou však spíše deskriptivního než algoritmického rázu a věnují se většinou pouze jednotlivým částem dané problematiky. Jediné možné řešení směrem k anotaci spíše anafory než koreference by bylo vyloučení z řady anaforů rematických uzlů (označených v tektogramatické anotaci jako f(ocus)), které častěji než jiné uzly vyjadřují koreferenci bez explicitně anaforického vztahu (srov. (10) a (12)b). Ve stávající anotaci jsme se však pro toto řešení nerozhodli, protože toto kritérium je příliš nepřesné – také rematické NP mohou odkazovat anaforicky[76] a naopak uzly v tématu často koreferují se svým antecedentem, aniž by na něj anaforicky odkazovaly (srov. (11)).

III.1.7. Princip zvláštní váhy podílu na kohezi textu

Při anotaci nejednoznačných případů bereme v úvahu, zda daný vztah přispívá ke kohezi textu. Pokud koreference jmenných frází není samozřejmá ani vztah mezi nimi nepřispívá ke kohezi textu, koreferenční vztah nemusí být označen. Stejná zásada platí pro asociační anaforu.

[74] Viz např. Poesio 2006.

[75] Srov. např. pro slovanské jazyky celkově Benacchio 1998, Birkenmaier 1979, Koseska-Toszewa 1983, Mendozová 2004, Nikolajevová 1979, Weiss 1983 aj.; pro češtinu Adamec 1980, Daneš 1999, Hlavsa 1972, Uhlířová 1996 aj.; pro ruštinu Bel'skij 1956, Boguslavskaja – Murav'evová 1987, Corbett 1986, Gladrow 1979, 1992, Golovačevová 1979, Padučevová 1988, Pospelov 1970, Šmelev 1984, Yokoyamová 2005 aj.

[76] Srov. příklad Fr. Štíchy *Když jste odešli s Oldou do toho kina, udělali jsme si s Lucy ta játra* (Štícha 1999).

III.1.8. Omezení počtu vztahů z jednoho uzlu / na jeden uzel

1. Z jednoho uzlu nemůže vycházet a do jednoho uzlu nemůže vcházet více než jedná šipka textové koreference (srov. také princip dodržování koreferenčního řetězce, III.1.2.). Dodržení daného pravidla při anotaci rozšířené koreference je logické a relativně jednoduché.[77] Problémy vznikají při kompilaci probíhající anotace rozšířené koreference s původní pronominální textovou koreferencí, kde odkazování z jednoho k více uzlům bylo v některých případech povoleno (viz bod 3. v III.8.).

V případě asociační anafory platí omezení, že z jednoho uzlu nemohou vycházet a do jednoho uzlu nemohou vcházet šipky více než jednoho typu. Pokud jde o vztahy jednoho typu, z jednoho uzlu může vycházet a do jednoho uzlu může vcházet více vztahů asociační anafory typů „množina – podmnožina" a „část – celek". Srov. v (13)f jmenná fráze „země podél jihozápadní a jižní hranice Ruska" se rozumí jako množina pro uvedené v předchozím kontextu podmnožiny *země bývalé sovětské střední Asie, Gruzie a Ázerbájdžán, Ukrajina* a *Arménie.* Srov. také obrázek č. 6 anaforické věty (13)f:

(13) *a. V zemích bývalé sovětské střední Asie ohrožoval další hegemonii Moskvy především tlak zdola.*

[... 3 věty ...]

b. V Gruzii a v Ázerbájdžánu se proti dalšímu vlivu Moskvy postavila jak veřejnost, tak vláda.

c. Obchody velmocí i obratně využitá vnitřní křehkost těchto států však ruským vojákům umožnily návrat.

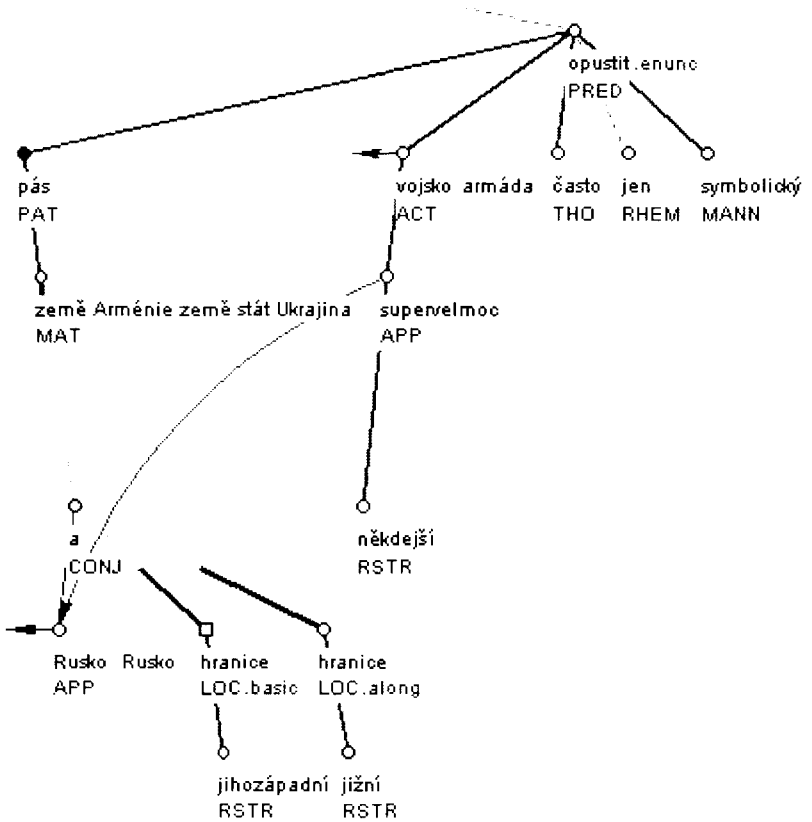
[... 7 vět ...]

d. Nový prezident Ukrajiny se prosadil pod heslem spolupráce s Ruskem.

e. Jen Bělorusku a Arménii se podařilo vyhnout větším sporům s Ruskem.

f. Pás zemí podél jihozápadní a jižní hranice Ruska opustila vojska někdejší supervelmoci často jen symbolicky.

[77] Jsou sice některé problémy technického rázu, kdy se zdvojují automaticky předgenerované a manuálně oanotované vztahy, tyto problémy se však průběžně řeší.



Obrázek č. 6: Několik šipek vztahu asociační anafory

2. Z jednoho uzlu může vycházet a do jednoho uzlu může vcházet jeden vztah identické textové koreference a zároveň jeden nebo více (v případě typů „množina – podmnožina" a „část – celek") vztahů asociační anafory.

3. Gramatická koreference nemá omezení počtu vztahů vycházejících z jednoho uzlu. V některých případech souvýskyt gramatické a textové koreference pak vyvolává technické problémy. (viz III.4.2.3.5.).

III.1.9. Preference anaforického vztahu před kataforickým

Při možnosti výběru mezi anaforickým a kataforickým odkazováním, odkazujeme anaforicky doleva.

Směr vztahu určujeme podle pořadí výrazu v povrchové struktuře věty, nikoliv podle pořadí uzlů v tektogramatickém stromě. Uzly ve stromě jsou totiž uspořádány podle jejich výpovědní dynamičnosti,[78] zatímco koherence textu a tedy i vnitrotextové odkazy jsou definovány lineárně na povrchu.

[78] Viz Mikulová a kol. (2005, s. 1060n.).

III.1.10. Princip jednofázové anotace

Na rozdíl od většiny schémat anotace koreference a anaforických vztahů, nedělíme proces anotace do dvou fází: vyhledávání potenciálních anaforických jednotek, kandidátů na anaforický vztah (tzv. „markables“) a anotaci komplexního vztahu. (Podrobněji k vydělování „markables“ viz v II.3.1.–II.3.7.). Nedostatkem tohoto řešení je, že výsledky naší anotace nejsou formálně srovnatelné s výsledky jiných anotací. Nicméně jsme se rozhodli nevyčleňovat předběžnou fázi anotace „markables“, a to z následujících důvodů: čeština nemá kategorii určenosti, nemáme tedy jednoduchý mechanismus vyčleňovaní potenciálně odkazujících jmenných frází (viz III.1.6.), pokud bychom se přesto rozhodli pro předběžnou fázi anotace „markables“, museli bychom stanovit jiná kritéria, než vyhledávání určených NP. Tato kritéria by byla buď formální (na základě gramatických informace o slovním druhu, rodu, pádu a čísle, pozice ve větě, aktuálního členění apod.) nebo sémantická a pragmatická. Gramatická informace je již obsažena v tektogramatické struktuře, vybíráme-li tedy předvolení „markables“ na základě formálních kritérií, můžeme tvrdit, že už je máme. Co se týče výběru „markables“ na základě sémantických a pragmatických kritérií, je to úkol natolik složitý, časově náročný a orientovaný na subjektivní názor anotátora, že mezianotátorská shoda této první fáze anotace koreference pravděpodobně nebude výrazně větší, než shoda celé stávající anotace rozšířené koreference a asociační anafory.

III.2. Formální charakteristika koreferovaných uzlů

V této části se chceme věnovat formální stránce výrazů, které se účastní koreferenčních vztahů na etapě anotace rozšířené koreference a asociační anafory na tektogramatické rovině PDT. Především obracíme pozornost k slovnědruhové informaci o koreferovaných párech a jejich detailnější gramatické charakteristice (III.2.1.–III.2.4.). Vycházíme z předpokladu, že se tato informace pro koreferenci a asociační anaforu bude podstatně lišit pouze frekvenčně, přehled slovnědruhových možností však zůstává stejný.

Při klasifikaci formy koreferovaných párů vycházíme především z formy anaforu (druhého členu páru). Vzhledem k tomu, že koreferenční vztah je symetrický (viz III.4.), antecedent se v ničím od anaforu neliší.[79] Výjimkou je koreference se situací (slovesem), která má jinou sémantickou interpretaci než tradiční koreference a vztah se nemůže považovat za symetrický (III.2.1.4.).

Jako kořen podstromu anaforu v tektogramatickém stromě mohou být:

- komplexní uzly[80] – autosémantické lexikální jednotky, aktuální elipsy, zájmena aj. (viz III.2.1.);
- kvazikomplexní uzly[81] – interpunkční znaménka, nealfanumerické symboly aj. (viz III.2.2.);
- kořeny souřadných struktur (viz III.2.3.);
- kořeny seznamových struktur[82] (viz III.2.4.).

III.2.1. Komplexní uzel v pozici anafora

Východiskem naší slovnědruhové analýzy je klasifikace sémantických slovních druhů a jejich vnitřní klasifikace, která je představena v Mikulové a kol. (2005, s. 37n.). Jde o seskupení do čtyř základních skupin, které jsou dále vnitřně členěny. Jsou to sémantická substantiva, adjektiva, adverbia a slovesa. Sémantické slovní druhy jsou kategoriemi tektogramatické roviny a odpovídají základním onomaziologickým kategoriím (substance,

[79] Anafor jako východisko pro klasifikace formálního vyjádření má také jinou intuitivní příčinu. Při postulování vztahu vycházíme od anaforu a hledáme pro něj antecedent v předcházejícím kontextu. Tato skutečnost pravděpodobně také částečně porušuje symetričnost koreferenčního vztahu v naší anotaci. Například, setkáme-li se s NP *Moskva a Budapešť*, vyhledáváme v předcházejícím kontextu tyto antecedenty spíše zvlášť než spojené koordinací.

[80] K pojmu „komplexní uzly" viz Mikulová a kol. (2005, s. 32).

[81] Ibid. 18.

[82] Ibid. 16.

vlastnost, okolnost, událost), což je pro naši práci velkou výhodou – na jedné straně se reference jazykových jednotek opírá právě o sémantické vlastnosti výrazů, na druhé straně informace o sémantických slovních druzích již zapracovaná do tektogramatické roviny pomůže při vyhledávání a klasifikaci koreferenčních vztahů a vztahů asociační anafory, příp. i pro jejich následnou analýzu.

Přináležitost komplexního uzlu k sémantickému slovnímu druhu je zaznamenána v atributu `sempos`.[83]

III.2.1.1. Sémantické substantivum v pozici anaforu

Pro anotaci rozšířené koreference a asociační anafory jsou substantiva nejfrekventovanějšími koreferujícími výrazy. Sémantická substantiva jsou dále vnitřně členěna, přičemž všechny podtypy sémantických substantiv se mohou zúčastnit koreferenčního vztahu. Sémantická substantiva tvoří následující podskupiny:

1. Pojmenovací sémantická substantiva (tradiční substantiva typu *otec, Marta*; posesivní adjektiva typu *otcův, Martin,* která mají t-lemma odpovídajícího substantiva). V tektogramatické struktuře má tento typ hodnotu atributu `sempos = n.denot`.

Srov. koreferenci přivlastňovacího adjektiva „podnikatelův" v (1)b:

(1) *a. Tímto faktorem je podnikatel – inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk, ani ztrátu.*
b. Podnikatelova {coref_text, na „podnikatel" v *a.*} *odměna, zisk, má však svůj původ nikoliv ve fungování, ale v rozbití stacionárního systému.*

2. Pojmenovací sémantická substantiva s odděleně reprezentovaným příznakem negace. Do této skupiny patří deverbativní substantiva zakončená na -ní / -tí (*hlasování*) a deadjektivní substantiva zakončená na -ost *(nezralost)*. V tektogramatické struktuře tento typ má význam atributu `sempos = n.denot.neg`.

3. Určitá pronominální sémantická substantiva ukazovací. Do této podskupiny patří ukazovací zájmena v pozici syntaktického substantiva (*Ten už nepřijde. O tohle mi nejde.*). Koreference většiny použití těchto zájmen již byla oanotována v původní verzi anotace

[83] Ibid. 33, 38n.

koreference (Kučová a kol. 2003, Mikulová a kol. 2005). V tektogramatické struktuře tento typ má význam atributu `sempos = n.pron.def.demon.`

4. Určitá pronominální sémantická substantiva osobní. Do této podskupiny patří všechna osobní zájmena a jejich posesivní protějšky (např. *já, můj*) včetně zájmen reflexivních, přičemž jde jak o uzly reprezentující povrchově realizovaná zájmena, tak o uzly nově vytvořené. V tektogramatické struktuře má tento typ význam atributu `sempos = n.pron.def.pers.`

Koreferenci osobních zájmen 3. osoby, včetně reflexivních zájmen (*se/si, svůj*), zpracovává původní anotace gramatické a textové koreference. Ve fázi anotace rozšířené koreference a asociační anafory osobní zájmena (uzly s t-lemmatem `#PersPron`) jsou anotovány hlavně jako antecedenty textově koreferenčních vztahů. Jde o prodlužování již existujících koreferenčních řetězců. Na obrázku č. 7 je zobrazena posloupnost koreferenčních NP (*Petr – on – Petr – on*). Původní zájmenná koreference zaznamenává dva koreferenční vztahy *Petr – on* (první a třetí šipka). Vztah *on – Petr* v původní anotaci zohledněn nebyl a zaznamenává se teprv ve stávající anotaci rozšířené koreference (druhá šipka).

Obrázek č. 7: Prodlužování existujících koreferenčních řetězců

Anotace koreference zájmen první a druhé osoby se ani v původní pronominální, ani ve stávající rozšířené identické a asociační anafoře systematicky neprovádí. Srov. např. nepropojené NP `#PersPron` *(ON)* – `#PersPron` *(JÁ)* – *Petr Chodura, podnikatel* – `#PersPron` *(VY)* v (2)a–d:

(2) *a.* `#PersPron.ACT` *Začal podnikat a vystřízlivěl [...]*

b. `#PersPron.ACT` {žádný koreferenční odkaz} *Byl jsem úplně naměkko, neschopen mluvit.*

c. Tak hodnotí Petr Chodura, podnikatel {žádný koreferenční odkaz} *z Ostravy, první momenty po oznámení, že se stal Vynikajícím podnikatelem roku 1993. [...]*
d. Takže #PersPron.ACT {žádný koreferenční odkaz} *jste se cítil schopen „jít do toho"?*

Naše rozhodnutí neanotovat koreferenci u zájmen první a druhé osoby je podmíněno následujícími příčinami:

- Specifikou textů korpusu PDT, ve kterých se dialogy vyskytují velmi zřídka.
- V různých dialogických replikách mezi osobními zájmeny první a druhé osoby a pojmenováním ve třetí osobě nemohou fungovat anaforická pravidla.
- V real-time dialozích, které se anotují na anglickém materiálu (Cinková a kol. 2009) se v tomto případě označuje exoforický odkaz na identifikační číslo dané entity. V PDT 2.0 to však není možné, protože rozhovory jsou zahrnuty do celkové struktury textu a zájmena první a druhé osoby odkazují nikoliv exoforicky ale endoforicky.
- Tento úkol je zřetelně vymezený vůči ostatním částem anotace a je možné jej realizovat později dodatečně, automaticky nebo částečně automaticky.
- Konvence a možná chyba – také není vyloučeno, že rozhodnutí neanotovat koreferenci u zájmen první a druhé osoby není zcela správné. Pravděpodobně by to nezapříčinilo žádné problémy, pouze nebylo včas odhaleno.

5. Neurčitá pronominální sémantická substantiva. Do této podskupiny patří vztažná zájmena *kdo, co, který / jenž* a *jaký*, plní-li funkci syntaktického substantiva (např. *Knihu, kterou si přál, nemohla sehnat,* nikoliv však *Kterou knihu si přál?*), jejich deriváty rovněž pouze v pozici syntaktického substantiva, tj. neurčitá zájmena (*někdo, některý)*, tázací (*kdo, který*), záporná (*nikdo*) a totalizační (*každý, všechen*). V tektogramatické struktuře tento typ má význam atributu `sempos = n.pron.indef`.

Vztažná zájmena byla již částečně zohledněna při anotaci gramatické a pronominální textové koreference a ve stávající anotaci působí jako antecedenty koreferenčních vztahů nebo jako členy vztahu asociační anafory (viz podobně u osobních zájmen v předcházejícím bodě č. 4.). Ostatní uvedená zájmena v původní anotaci koreference zpracována nebyla a anotujeme je ve stávající anotaci. Srov. (3):

(3) *a. X daroval Y počítače, kopírky apod.*

b. Vše {bridging na "počítače", "kopírky" v *a.*} *v hodnotě 1 milión.*

6. Číslovky v pozici syntaktického substantiva (např. *tři* ve větě *Vybrali tři* nikoliv však v *tři knihy*, kde *tři* je považováno za sémantické adjektivum) a dílové číslovky typu *třetina.* Srov. (4)a–b:

(4) *a. Připomenu, že po vstupu vojsk SSSR 21. srpna 1968 na naše území jsem se zdržel (vraceje se z jugoslávských prázdnin) ve Vídni.*

b. Vzpomínám na takzvané zelené hranice zcela bezbariérové a na dosud nevídanou blahovůli zahraničních a našich celních a policejních orgánů už na jaře 1968 {coref_text na „1968"}...

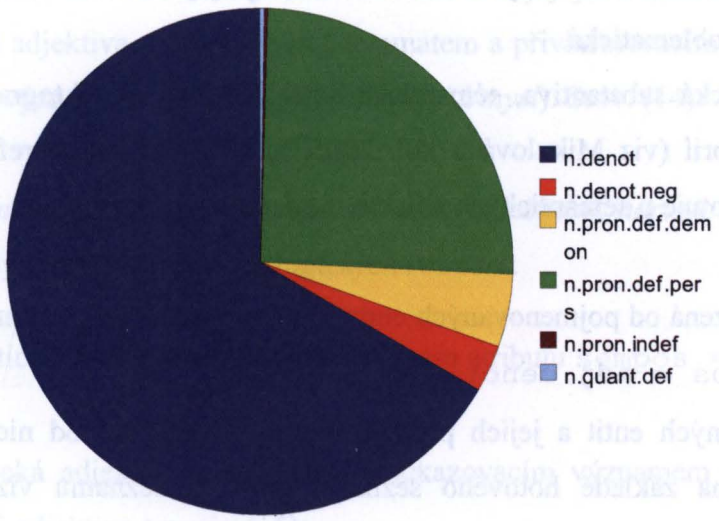
V tektogramatické struktuře tento typ má význam atributu `sempos = n.quant.def`.

Následující tabulka č. 9. představuje shrnutí výskytů sémantických substantiv v pozici anaforu:

sempos	počet	%	příklad
n.denot	11408	66,6	*(5) a. Podle ČSBS nemůže zákon plně nahradit škody, a konečným plátcem musí být proto Německo.* *b. Preambule zákona prý vládu dostatečně zavazuje k tomu, aby toto odškodnění na Německu vymohla.*
n.denot.neg	515	3	*(6) Deset dní před fotbalovým derby Sparty se Slavií není jasné, kde se bude toto ligové utkání hrát.*
n.pron.def.demon	805	4,7	*(7) K tomu byla ustanovena pracovní skupina, která se tímto bude zabývat.*
n.pron.def.pers	4330	25	viz (2)
n.pron.indef	36	0,2	*(8) Z téměř tří desítek smluv upravujících vztahy mezi oběma subjekty celního soustátí jsou okamžitě vypověditelné všechny...* *(9) Všichni to věděli a všichni na něho byli krátcí, protože Franta Kejval měl svá lidská a občanská práva.*
n.quant.def	51	0,3	viz (4)a–c.
CELKEM	17135		–

Tabulka č. 9: Sémantická substantiva v pozici anaforu

Celkové rozložení jednotlivých typů sémantických substantiv v pozici anaforu koreferenčního vztahu ukazuje graf č. 1. Ačkoliv v dané pozici byly zastoupeny všechny typy, existuje v jejich výskytu výrazný nepoměr.



Graf č. 1: Rozložení jednotlivých typů sémantických substantiv v pozici anaforu textově koreferenčnícho vztahu

III.2.1.2. Sémantické adjektivum v pozici anaforu

Primární funkcí prototypického adjektiva je predikace, nikoliv denotace. Přesto některá adjektiva plní v určitých kontextech především denotační funkci. V těchto případech jsme se rozhodli pro anotaci koreference a asociační anafory u adjektiv.

Rozhodnutí pro anotaci koreference u některých typů adjektiv je podmíněno ještě jednou, spíše aplikační příčinou. Jedním z cílů anotace koreference na PDT je možnost jejího využití pro strojový překlad především do angličtiny. V angličtině však je gramatická kategorie adjektiv blízká substantivům: tyto dva slovní druhy se často liší pouze syntaktickou pozici a funkcí ve větě. Srov. např. *Prague (Praha – pražský)* v adjektivní a substantivní funkci v češtině a angličtině:

(10) angl. *He arrived in Prague and found the Prague atmosphere quite casual.*
č. *Přijel do Prahy a pražská atmosféra se mu zdála celkem neformální.*

Zpracování koreference této kategorie adjektiv (*prenominal modifiers, bare nouns*) jsou věnovány celé kapitoly manuálů anotace koreference a většinou je alespoň část těchto adjektiv anotována na koreferenci stejně jako referenční substantiva (viz např. MATE: Poesio 2004;

MUC: Hischman 1997; PoCoS: Chiarchos – Krasavina 2007 aj.). Z toho důvodu je užitečné mít anotaci koreference referenčních adjektiv v češtině, i když jejich klasifikace na referenční a nereferenční je často problematická.

Stejně jako sémantická substantiva, sémantická adjektiva jsou na tektogramatické rovině dále členitelnou kategorií (viz Mikulová a kol. 2005, s. 67). Textová koreference a vztah asociační anafory se anotuje u sémantických adjektiv následujících podskupin:

1. Adjektiva odvozená od pojmenovaných entit. V tektogramatické struktuře tento typ má význam atributu `sempos = adj.denot`.

Anotace pojmenovaných entit a jejich propojování s odvozenými od nich adjektivy se provádí automaticky na základě hotového seznamu párů (k seznamu viz podrobněji v IV.1.2.1.).

Následující příklady (11)–(13) znázorňují koreferenci adjektiv odvozených od pojmenovaných entit:

(11) *a. Radní jednomyslně vyjádřili nesouhlas s přítomností chovanců společnosti Struktura v Košťanech.*
b. Ředitelka košťanské {coref_text na „Košťany"} *základní školy Jarmila Hejduková byla jednou z iniciátorek podpisové akce.*
c. Pikantní detail v celé záležitosti je, že třinácti až čtrnáctiletí chlapci si dům v Košťanech {coref_text na „košťanský"} *teprve upravovali.*

(12) *a. Významnou roli v dějinách žvýkačky sehrál mexický diktátor Antonio Lopez de Santa Anna.*
b. Poté , co byl v roce 1845 jako prezident svržen a na deset let vypovězen na Kubu , vydal se do New Yorku s jedinou myšlenkou – získat zpět vládu nad Mexikem {coref_text na „mexický"}.

(13) *a. Když se opat oseckého kláštera dověděl, jaké problémy s nimi jsou, další pobyt zakázal a němečtí hoši se museli i s vychovatelem z kláštera vystěhovat.*
b. Někteří se vrátili do Německa {coref_text na „německý"}, *další přešli na faru v nedalekém Jeníkově a spojili se se skupinou, která tady byla již ubytována.*

2. Sémantická adjektiva s přivlastňovacím významem. Posesivní adjektiva s t-lemmatem odpovídajícího substantiva se anotují na koreferenci a byly již rozebrány v III.2.1.1. Do této podskupiny patří adjektiva s adjektivním t-lemmatem a přivlastňovacím významem (*palácový – palác, dětský* s významem *'dítěte'* např. v NP *dětská mysl*). Srov. (14):

(14) *Co se může dospělému zdát zanedbatelnou záležitostí, naroste v dětské* {bridging na „dospělý"} *mysli třeba i do tragických rozměrů.*

V tektogramatické struktuře tento typ má význam atributu `sempos = adj.denot`.

3. Sémantická adjektiva s ukazovacím/odkazovacím významem (*tamní, zdejší* apod.). Srov. koreferenci adjektiva *tamní* v (15):

(15) *Jakkoliv mé vystoupení ve Würzburku obestírala krásná a klidná pohoda, tak o rok pozdější mou přednášku v Heidelberku ve velké posluchárně tamní* {coref_text na „Heidelberk"} *proslulé univerzity, k níž patřil i Institut pro studium slovanských jazyků a literatur, provázel protest, napadení skupinou posluchačů a jejich mluvčí vystoupil proti mně velmi adresně, a i když více méně zaobaleně, i hrubě.*

V tektogramatické struktuře tento typ má význam atributu `sempos = adj.denot`.

4. Jiná referující sémantická adjektiva. Srov. např. *podnikatelský* v (16)b:

(16) *a. Podnikatelova odměna, zisk, má však svůj původ nikoliv ve fungování, ale v rozbití stacionárního systému.*
b. Tento druh podnikatelské {coref_text na „podnikatelův"} *odměny je vlastně monopolní rentou a je dočasné povahy.*

V tektogramatické struktuře tento typ má význam atributu `sempos = adj.denot`.

5. Základní určité číslovky vystupující v pozici syntaktického adjektiva se mohou zúčastnit koreferenčního vztahu, jsou-li antecedentem číslovky v pozici syntaktického substantiva (viz (4)).

Neanotujeme následující podskupiny sémantických adjektiv:

1. Pojmenovací sémantická adjektiva (`sempos = adj.denot`), kromě podskupin uvedených výše v 1–5, tj. „běžná" adjektiva jako *červený, vlastní* aj.
2. Ukazovací a identifikační zájmena v pozici syntaktického adjektiva (*Ten dům už koupili. Takový přístup se mi nelíbí.*). V tektogramatické struktuře tento typ má význam atributu `sempos = adj.pron.def.demon`.
3. Vztažná zájmena *který* a *jaký*, jsou-li v pozici syntaktického adjektiva, a jejich deriváty rovněž pouze v pozici syntaktického adjektiva (*Kup mu nějakou knihu. Každý člověk má problémy.*) V tektogramatické struktuře tento typ má význam atributu `sempos = adj.pron.indef`.
4. Číslovky v adjektivní syntaktické pozici kromě případů popsaných v bodě 5 výše.

Následující tabulka č. 10. představuje shrnutí výskytů sémantických adjektiv v pozici anaforu:

typ	počet	příklad
adj.denot od NE[84]	–*	viz (11)–(13)
adj.denot POSS	–*	viz (14)
adj.denot DEMON	–*	viz (15)
adj.denot jiná	–*	viz (16)
číslovky	–*	viz (4)
CELKEM adj.denot	1130	–

* Nelze spočítat, protože tektogramatické rovina neobsahuje odpovídající atribut.

Tabulka č. 10: Sémantická adjektiva v pozici anaforu

III.2.1.3. Sémantické adverbium v pozici anaforu

Podobně jako je tomu u adjektiv (viz III.2.1.2.), je primární funkcí prototypického adverbia predikace, nikoliv denotace. Považujeme však za smysluplné anotovat koreferenci a asociační anaforu u určitých pronominálních sémantických adverbií ukazovacích a identifikačních (např.: *tady, tam(hle), tehdy, tak, tenkrát* a *tamtéž*) a u derivátů těchto adverbií (např. *tudy* a *zde* od

[84] Zde a dále NE = named entity = pojmenovaná entita.

tady). Textovou koreferenci neanotujeme u ostatních sémantických adverbií *(teď , potom, předtím, proto, jakž).* V tektogramatické struktuře má tento typ význam atributu `sempos = adv.pron.def`.

Srov. (17)a–b:

(17) *a. Potvrdil to mimo jiné průzkum, který uskutečnilo ministerstvo zemědělství společně s ministerstvem zdravotnictví v okrese Plzeň-jih.*
b. Normě tu {coref_text na „okres"} *neodpovídalo 97.6 procenta odebraných vzorků.*

Adverbia se neanotují, pokud mají tektogramatický funktor PREC. Toto omezení má dva důvody: za prvé, od uzlů s funktorem PREC se neočekává žádná reference (je to atomický uzel); za druhé, funktor PREC je definován na klauzích jako uzel, který „reprezentuje výraz signalizující návaznost klauze na předcházející kontext" (Mikulová 2005, s. 534), tedy anotace textových vztahů naznačených jazykovými jednotkami s funktorem PREC patří spíše do anotace diskurzu, která právě probíhá na PDT (Mladová a kol. 2008). Srov. (18)a–b:

(18) *a. Jsem si jist, že ne hony na čarodějnice, ale právě poukázání a následné právní kroky, učiněné vůči opravdovým pachatelům násilí a komunistické svévole, je povzbudivým signálem toho, že naše společnost sice pozdě, ale přece jen vyvodila praktický a konkrétní krok, potvrzující přesvědčení, že historii tvoří konkrétní lidé, kteří ve svých činech projevují svou svobodnou vůli a nesou tedy za ně i svou osobní odpovědnost.*
b. Ukazuje se tak {žádný koreferenční vztah}, *že pokud společnost chce, může najít onu pomyslnou dělicí čáru.*

Následující tabulka č. 11 shrnuje nejčastější výskyty sémantických adverbií v pozici anaforu:

t-lemma	počet	příklad
tady	82 (z 158[85])	viz (17)a–b.
tam	36 (z 56)	*(19)a. Více než deset let se v Belfastu pravidelně schází skupina lidí všech vyznání v katedrále sv. Anny.* *b. Každé poledne se tam* {coref_text na „katedrála"} *společně pomodlí za mír v Severním Irsku.*
tehdy	11 (z 25)	*(20) a. Teprve pod vedením trenéra Sammyho Leeho, bývalého olympijského vítěze, se dostal Louganis na výsluní sportovní slávy: v Montrealu získal r. 1976 svou první olympijskou medaili – stříbrnou.* *b. Tehdy* {coref_text na „rok 1976"} *mu bylo 16 let.*
CELKEM	129	–

Tabulka č. 11: Sémantická adverbia v pozici anaforu

III.2.1.4. Sémantické sloveso jako člen koreferenčního vztahu

Ze všech anotovaných slovních druhů má sloveso nejmenší, ba žádný referenční potenciál. Avšak sloveso (slovesná fráze, klauze, věta se slovesem v kořenu, celá situace popsaná více než jednou větou) může být antecedentem anaforické jmenné fráze a v tom případě podléhá anotaci koreference. Srov. (21):

(21) *a. Jistotu v tomto směru dávají nejnovější kroky vlády SR, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience.*
b. Byť má na tento krok {coref_text na „zavést"} *určité právo, v daném okamžiku však vyznívá jako tvrdé politické rozhodnutí vlády, která se snaží velice rezolutními administrativními kroky zredukovat mnohamilionové pasívum v obchodní výměně s ČR.*

V některých případech kontext láká k anotaci koreference mezi slovesy nebo mezi slovesem v pozici anaforického členu a jmennou frází v pozici antecedentu, proto anotátoři tento vztah chybně zaznamenávají jako koreferenční. Srov. např. vztah mezi *převýchova* a *převychovávat* v

[85] Počet uvedený v závorce je celkový počet výrazů s daným lemmatem v PDT včetně neanotovaných na koreferenci.

(22)a–b a část řetězce generických použití *žvýkačky – nežvýkají – žvýkání – žvýkali – žvýkání* v (23)a–e.

(22) *a. Na převýchovu se pokud vím, posílali ti, kteří měli podle těchto zrůdných režimů nevhodný původ.*
b. Naše sdružení nepřevychovává, ale snaží se vychovávat.

(23) *a. Vybrané kapitoly z dějin žvýkačky.*
b. Lidi nežvýkají, to jenom krávy.
c. Pravda o tom, že žvýkání pro žvýkání bylo odjakživa činností veskrze lidskou – kam paměť lidského rodu sahá.
d. Se stejnou radostí však zamlčí, že Řekové často a s oblibou žvýkali kousky ztuhlé mízy mastikového keře, který se pěstuje především na ostrově Chios.
e. Známý antický lékař a botanik Dioscorides psal v prvním století našeho letopočtu obsáhle o léčebném a hygienickém účinku žvýkání.

Ani koreferenci mezi slovesy, ani mezi slovesem v pozici anafora a jmennou frázi v pozici antecedentu však v anotaci nezaznamenáváme. Nečiníme to z několika důvodů. Za prvé, slovesa nemají referenci v tradičním slova smyslu a tedy nemohou být mezi sebou koreferenční ani na sebe odkazovat. Za druhé, systematické zapojení sloves do koreferenčních řetězců nás přinutí anotovat koreferenci mezi všemi stejnými slovesy odkazující na stejnou situaci (např. *A řekl, že B – Řekl to včera*), a to už také vychází za rámce koreference a anaforického odkazování.

V anotovaném korpusu se však našly 34 příklady, kde anotátoři chybně zaznamenali koreferenční vztah vedoucí od slovesa. Srov. (24)a–b, (25)a–b a jiné.

(24) *a. Jeho úřad 116 projektů předal opět Luxovu ministerstvu zemědělství k aktualizaci.*
b. Ministerstvo zemědělství stačilo zaktualizovat 84 projektů, které pak znovu převzali lidé ministra Skalického.

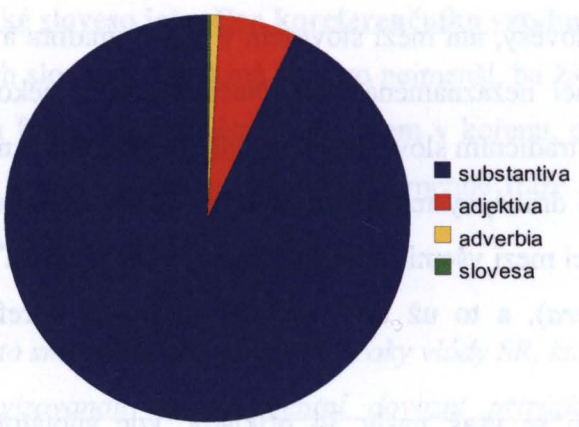
(25) *a. V každém případě vláda má vlastní dlouhodobou koncepci daňové politiky založenou na snižování daňového břemene.*
b. Záměr postupně snižovat daně hlásá jak vláda, tak i podnikatelé.

* * *

Celkově k použití komplexních uzlů v koreferenčních vztazích a ve vztazích asociační anafory můžeme shrnout, že jejich počet ubývá směrem od sémantických substantiv ke slovesům. Rozložení komplexních uzlů v pozici anaforu koreferenčního vztahu znázorňuje tabulka č. 12 a graf č. 2.

sémantický slovní druh	počet výskytů	% k ostatním typům
sémantická substantiva	17135	93
sémantická adjektiva	1130	6
sémantická adverbia	129	0,7
sémantická slovesa	58	0,3

Tabulka č. 12: Komplexní uzly v pozici anaforu koreferenčního vztahu



Graf č. 2: Rozložení komplexních uzlů v pozici anaforu koreferenčního vztahu

III.2.2. Kvazikomplexní uzel v pozici anaforu

Kvazikomplexní uzly jsou ty, které v tektogramatickém stromu vystupují ve stejných pozicích jako komplexní, ale nemají gramatické charakteristiky. Jde buď o nově vytvořené uzly v pozicích uzlů nesoucích valenční a nevalenční doplnění, nebo o uzly reprezentující povrchově přítomná interpunkční znaménka a jiné nealfanumerické symboly. Koreference nově vytvořených uzlů ve funkci valenčních a nevalenčních doplnění byla již oanotována v

etapě anotace gramatické a pronominální koreference (viz #Unsp v Kučová a kol. 2003, Mikulová a kol. 2005).

Ostatní případy kvazikomplexních uzlů jsou představeny v následující tabulce č. 13.

t-lemma	počet	příklad
#Percnt	6	*(26) a. Číslo týdne.* *b. 90 %* {coref_text na „číslo"}.

Tabulka č. 13: Kvazikomplexní uzly v pozici anaforu

III.2.3. Kořeny souřadných struktur v pozici anaforu

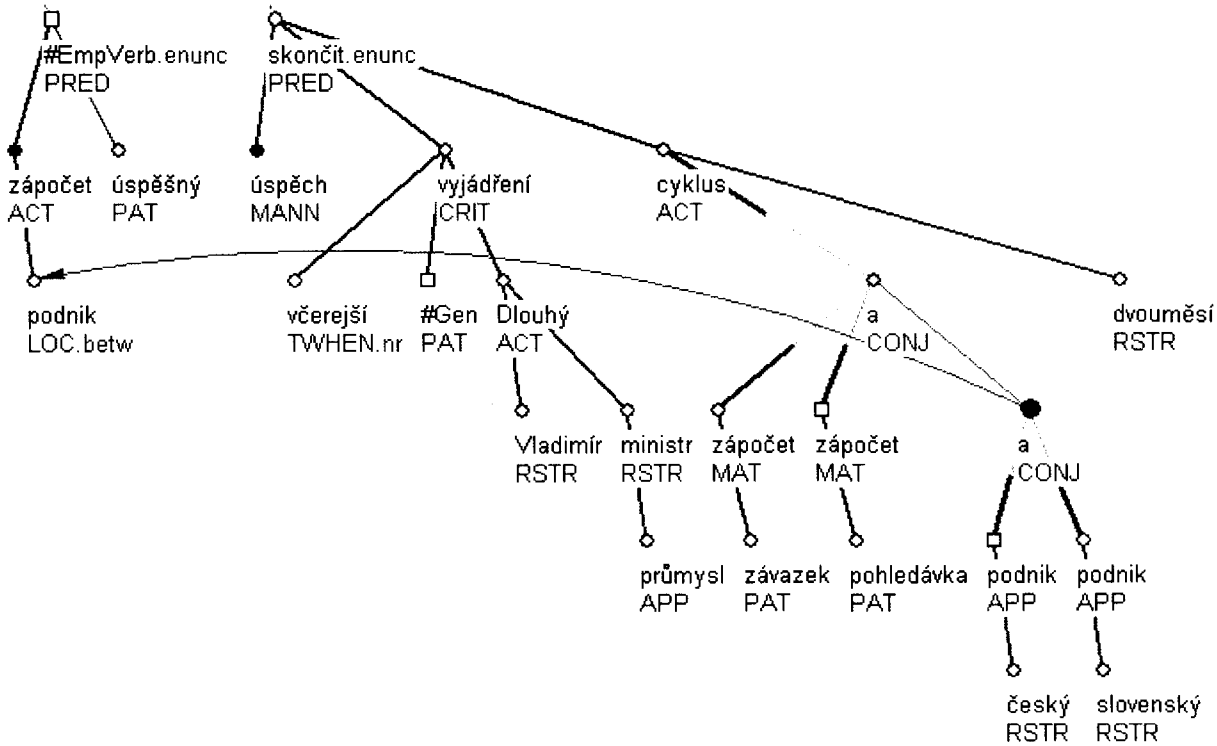
Kořeny souřadných struktur mohou být koordinační a apoziční spojky (*a, ale* aj.), zástupná t-lemmata pro interpunkční znaménka (např. #Comma, #Dash, #Colon aj.) nebo operátory matematických operací a intervalů (např. +, krát, od_ do).[86]

Kořeny souřadných struktur vystupují v pozici anaforu zcela běžně, jak v případě textové koreference, tak i v případě asociační anafory. Srov. např. (27) a obrázek č. 8 pro koreferenci a (28) pro asociační anaforu.

(27) *a. Zápočty mezi podniky úspěšné.*
b. Úspěchem skončil podle včerejšího vyjádření ministra průmyslu Vladimíra Dlouhého dvouměsíční cyklus zápočtů závazků a pohledávek českých a {coref_text na "podnik" v (27)a} *slovenských podniků.*

(28) *Celková suma započtených závazků a pohledávek činila 28.6 miliardy korun, z toho v České republice byly započteny pohledávky a* {bridging SET_SUB na "započtené závazky a pohledávky"} *závazky za 20.5 miliardy korun a na Slovensku za 8.1 miliardy korun.*

[86] Vysvětlení specifických názvů t-lemmat viz v Mikulová a kol. 2005.



Obrázek č. 8: Kořeny souřadných struktur v pozici anaforu

Následující tabulka č. 14 představuje shrnutí nejčastějších výskytů kořenů souřadných struktur v pozici anaforu a jejich celkový počet:

t-lemma	počet	příklad
a	61	viz (27)a–b.
#Comma	31	*(29) a. Dům za tři miliony do Liberce* *b. První cenu soutěže LN Zápisník akcionáře* ,#Comma {coref_text na "dům"} *luxusní domek v hodnotě 3000000 korun, vyhrál invalidní důchodce ing. Vladimír Duda z Liberce.*
#Dash	25	*(30) a. Americká pomoc parlamentu ohrožena spekulacemi* *b. Celou počítačovou síť, telefonní ústřednu, velice výkonné kopírovací stroje –.* #Dash {coref_text na "pomoc"} *to vše poskytly USA, resp. americký kongres československému federálnímu parlamentu v rámci pomoci parlamentům zemí bývalého sovětského bloku.*
#Bracket	16	*(31) a. Dráhy nabízejí zajímavá místa* *b. Přes rok existuje u Českých drah (.*#Bracket {coref_text na "Drahy"} *ČD) Divize majetkového podnikání a privatizace, jejíž hlavní náplní je pronájem, prodej a privatizace nemovitého majetku.*
CELKEM u kořenů souřadnýc h struktur	150	–

Tabulka č. 14: Kořeny souřadných struktur v pozici anaforu

III.2.4. Kořeny seznamových struktur v pozici anaforu

Kořeny seznamových struktur jsou nově vytvořené uzly tektogramatického stromu, které vystupují buď jako kořen struktury, která plní funkci názvu, např. knihy, skladby apod. (t-lemma #Idph) nebo shromažďuje do seznamu členy cizojazyčného výrazu (t-lemma #Forn). V případě odkazování na objekty, jejichž pojmenování jsou v tektogramatickém stromě zachycena jako seznamové struktury, se koreferenčního vztahu vždy zúčastní pouze řídící uzel toho seznamu, čili uzel s t-lemmatem #Idph nebo #Forn. Srov. (32)a–b pro #Idph a (33)a–c pro #Forn.

(32) *a. První stěžejní prací dr. Svobody a jeho spolupracovníků byl reléový počítač SAPO.*

b. Pracovali na něm od roku 1950, do zkušebního provozu byl však #Idph.PAT {coref_text na "na něm"} *SAPO.*ID *uveden – vzhledem k obtížím při získávání součástek i k informační bariéře – až roku 1957.*

(33) *a.* #Forn.ACT *TTI.*FPHR *Therm.*FPHR *dodával stále vodoměry nedělené.*

b. Firmě dlouho trvalo, než #PersPron.ACT *prosadila u německého producenta dělení vodoměrů.*

c. Jakmile #Forn.ACT {coref_text na #PersPron v (33)b} *TTI.*FPHR *Therm.*FPHR *začala dodávat dělené vodoměry, opět získala dřívější pozice na trhu.*

Používání kořenů seznamových struktur v pozici anaforu shrnujeme v následující tabulce č. 15:

t-lemma	počet	příklad
#Idph	29	viz (32)
#Forn	64	viz (33)
CELKEM	93	–

Tabulka č. 15: Kořeny seznamových struktur v pozici anaforu

III.3. Gramatická koreference

Gramatická koreference je takový typ koreference, kdy je možné určit antecedent na základě gramatických pravidel daného jazyka. V programu zobrazení tektogramatických stromů TrEd[87] gramatická koreference je zobrazena červenou šipkou, která vede od anaforu na antecedent.

Za gramatickou koreferenci se považuje

- koreference zvratných zájmen. Srov. (1), kde zvratné zájmeno *sobě* koreferuje se subjektem *matka*.

(1) *Sobě nedopřeje matka nikdy nic.*[88]

- koreference vztažných prostředků (*který, jenž* apod.). Srov. (2):

(2) *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.*

- koreference v recipročních konstrukcích. Od nově vytvořeného uzlu pro valenční doplnění chybějícího v důsledku reciprocity vede vždy vztah gramatické koreference k uzlu pro tu valenční pozici, která je s pozicí zastoupenou uzlem s t-lematem `#Rcp` ve vztahu reciprocity. Srov. (3), kde koreferenční vztah vede od nově vytvořeného uzlu s t-lemmatem `#Rcp` na subjekt *sultáni*.

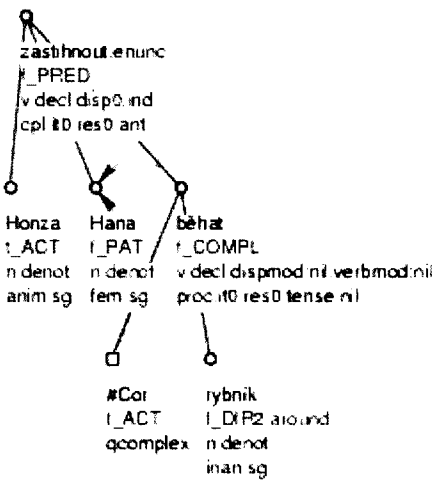
(3) *Sultáni se vystřídali* `{#Rcp.PAT}` *na trůnu.*

- koreference u doplnění s dvojí závislostí vyjádřených slovesnou formou. Ve vztahu gramatické koreference je jeden z povrchově nevyjádřených aktantů doplnění realizovaného slovesnou formou, které má tzv. dvojí závislost (trpné participium, přechodník, infinitiv v pozici doplňku (`COMPL`) nebo aktantu `EFF`(ektu)). Srov. koreferenci nevyjádřeného aktora infinitivu *běhat* s patientem *(Hanka)* řídícího slovesa *(zastihl)* v (4) a obrázku č. 9:

[87] Podrobněji o tomto programu viz dále v IV.1.1. a Pajas – Štěpánek 2008.

[88] Příklady (1)–(6) jsou převzaty z Mikulové a kol. (2005: 936n.).

(4) *Honza zastihl Hanku běhat kolem rybníka.*

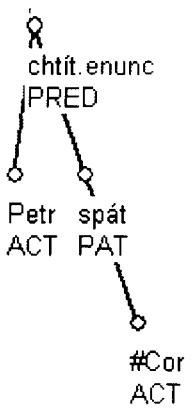


Obrázek č. 9: Gramatická koreference

- kontrola – nastává u určitých sloves (slovesa kontroly, např. *začít, dovolit, chtít, dokázat* apod.), která jsou zadána seznamem v manuálu k anotaci tektogramatické roviny PDT (Mikulová a kol. 2005). Jde o obligatorní nebo fakultativní koreferenční vztah mezi kontrolujícím a kontrolovaným členem, přičemž v zásadě platí, že
  i. kontrolující člen je jedním z členů valenčního rámce řídícího slovesa kontroly (např. v (5) *Petr* je člen valenčního rámce (aktor) řídícího slovesa kontroly *chtít*);
  ii. kontrolovaný člen je jedním z členů valenčního rámce infinitivu (nebo deverbativního substantiva) závislého na slovese kontroly (např. v (5) kontrolovaný člen `#Cor` je člen valenčního rámce (aktor) infinitivu *spát* závislého na slovese kontroly *chtít*);
  iii. infinitiv (nebo deverbativní substantivum), jehož valenční doplnění je v pozici kontrolovaného členu, je valenčním doplněním řídícího slovesa kontroly (např. v (5) infinitiv *spát* je valenčním doplněním (patientem) řídícího slovesa kontroly *chtít*).

2. Srov. koreferenční vztah mezi nevyjádřeným aktorem infinitivu *spát* a aktorem řídícího slovesa *chtít* v (5).

(5) *Petr chce spát.*

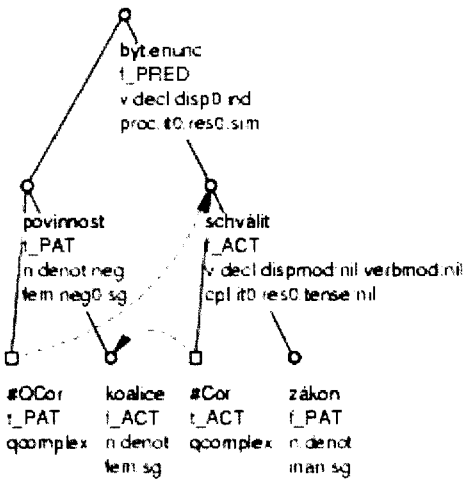


Obrázek č. 10: Gramatická koreference, kontrola

- kvazikontrola – specifický gramatický koreferenční vztah, který nastává u víceslovných predikátů, jejichž závislou část představuje nějaké substantivum, které má valenci. Srov. (6) a obrázek č. 11, kde závislou část víceslovního predikátu *je povinností koalice* představuje valenční substantivum *povinnost*, jehož aktor je ve vztahu gramatické koreference s aktorem řídícího slovesa:

(6) *Povinností koalice je schválit zákon.*[89]

[89] Příklad a obrázek jsou převzaty z Mikulové a kol. (2005, s. 1033).



Obrázek č. 11: Gramatická koreference – kvazikontrola

Podrobněji ke gramatické koreferenci viz v Mikulové a kol. 2005, s. 935n.

Anotace gramatické koreference byla provedena částečně automaticky na celém korpusu PDT 2.0. Výsledky automatických procedur a další informace jsou dokumentovány v (Kučová a kol. 2003). V anotaci rozšířené koreference zůstávají všechny původní šipky – do gramatické koreference jsme nezasahovali.

III.4. Textová koreference

Textovou koreferenci chápeme jako užití různých jazykových prostředků pro označení stejného objektu mimojazykové skutečnosti. Základním principem textové koreference je identita referentů antecedentu a anaforu.

Vztah koreference je

- symetrický (pokud A je koreferenční s B, B je koreferenční s A) a
- tranzitivní (pokud A je koreferenční s B a B je koreferenční s C, pak A je koreferenční s C).

Textové koreference se mohou zúčastnit výrazy v asertivních, otázkových, rozkazovacích i negovaných větách.

V současné fázi anotace rozlišujeme původní pronominální textovou koreferenci (stručný přehled viz v III.4.1., podrobný popis viz v (Kučová a kol. 2003, Mikulová a kol. 2005)) a rozšířenou textovou koreferenci (III.4.2.).

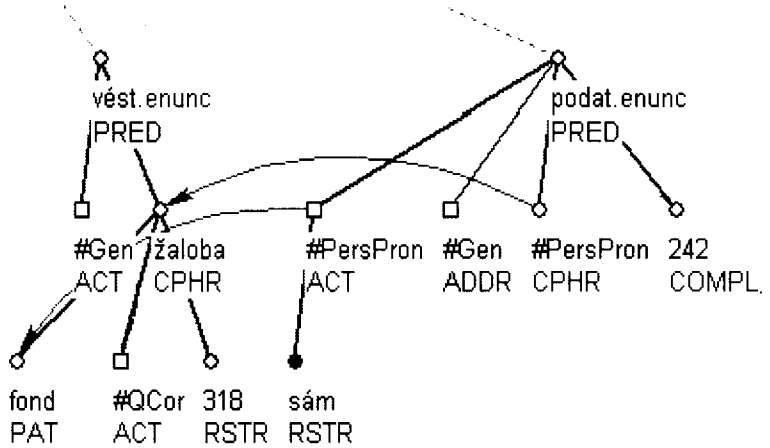
III.4.1. Pronominální textová koreference

Původní pronominální textová koreference je anotována ručně na celém korpusu PDT[90] a týká se většiny případů pronominalizace a elips. Při anotaci pronominální koreference se vyznačovaly následující vztahy:

1. Textová koreference u osobních a přivlastňovacích zájmen pro 3. osobu[91] (kromě reflexivních zájmen, která byla zpracována v rámci gramatické koreference, viz. III.3.). Tyto zájmena mají v tektogramatickém stromě jednotné t-lemma `#PersPron`. Srov. vztah mezi *jich* a *žaloba* v (1) a na obrázku č. 12:

[90] K anotaci pronominální koreference viz podrobněji (Mikulová a kol. 2005, Kučová a kol. 2003).

[91] Zájmena 1. a 2. osoby se nevyznačovala (viz vysvětlení v III.2.1.1.).



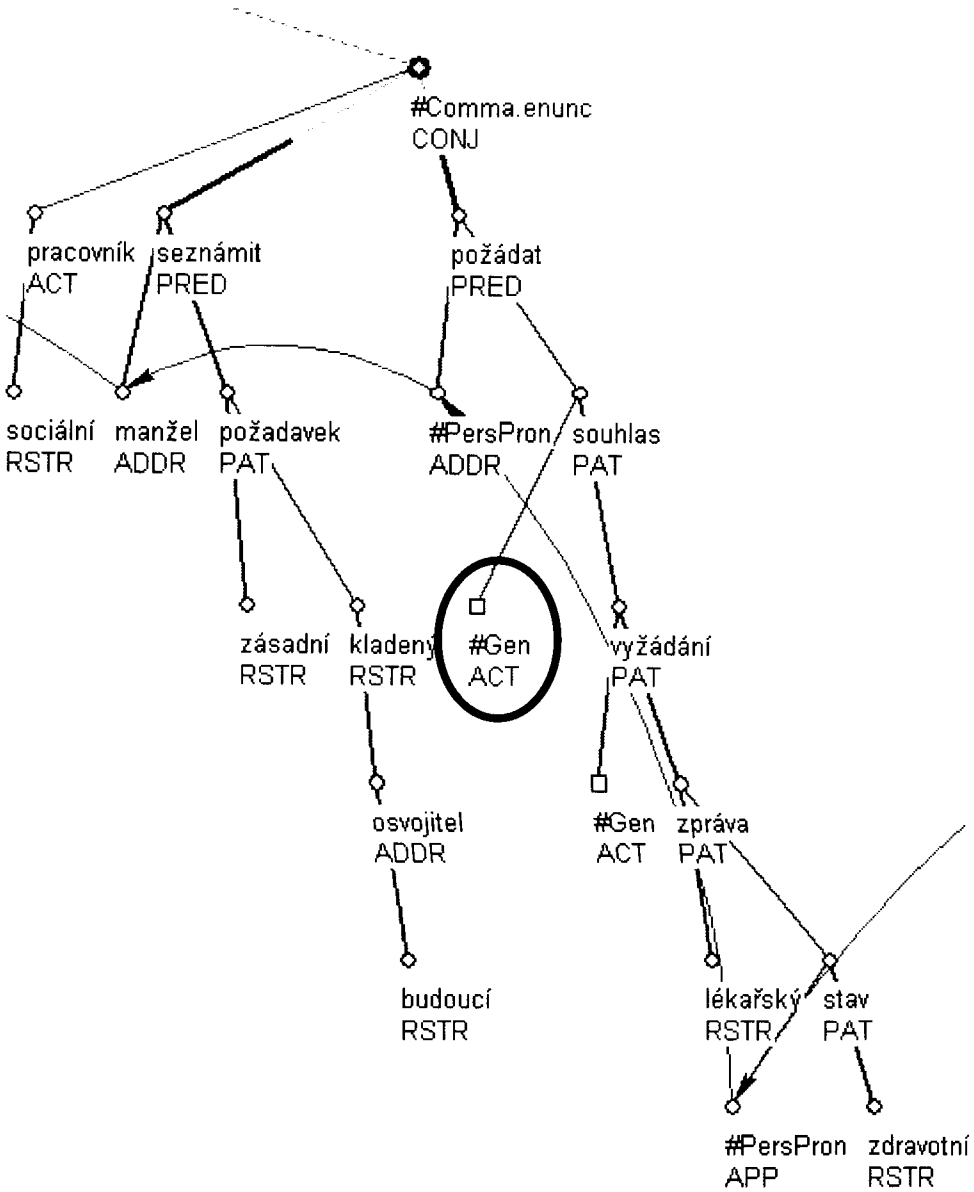
Obrázek č. 12: Textová pronominální koreference

(1) *a. Proti fondu je vedeno 318 žalob.*

b. Sám #PersPron {coref_text na „fond"} *jich* {coref_text na „žaloba"} *podal 242.*

2. Textová koreference u ukazovacích zájmen *ten, ta, to* v substantivní funkci.
3. Textová koreference při aktuální elipse, kdy je do tektogramatického stromu doplněn nový uzel se zástupným t-lematem `#PersPron`. Srov. koreferenci nově doplněného uzlu `#PersPron` a antecedentu *fond* v (1). Při doplňování závislých valenčních doplnění všeobecným aktantem v podobě t-lematu `#Gen` případná koreference u těchto uzlů nebyla zachycena. Srov. např. koreferenční řetězec *manžele – je – jejich* v (2) a obrázek č. 13:

(2) *Sociální pracovník seznámí manžele se zásadními požadavky kladenými na budoucí osvojitele, požádá je o* `#Gen` {žádný koreferenční odkaz} *souhlas k vyžádání lékařských zpráv o jejich zdravotním stavu.*



Obrázek č. 13: Textová pronominální koreference

Při anotaci původní zájmenné koreference se textová koreference chápala jako užití různých jazykových prostředků (v daném případě zájmen a elips), které **anaforicky** (zřídka kataforicky) **odkazují**, tj. převažovala orientace na odkazovací (anaforické) funkce členů, nikoliv na jejich referenční vlastnosti. Termín *koreference* se přitom používal spíše jako synonymum pro anaforické odkazování, nikoliv ve smyslu identity referentů antecedentu a anaforu. Z toho důvodu se v rámci textové koreference analyzovaly také jiné anaforické vztahy než koreferenční. V rámci pronominální textové koreference se rozlišovaly tři základní typy odkazování:

1. Odkazování k jednoznačnému, explicitnímu antecedentu – běžná pronominální koreference. Ve většině případu jde o skutečnou koreferenci. K výjimkám viz III.8. bod 2.
2. Odkazování k většímu úseku textu (více než jedna věta).
3. Exoforické odkazování (odkazování k mimotextové situaci či skutečnosti).

Všechny typy odkazování jsou přítomné i při anotaci rozšířené textové koreference. Odkazování k segmentu textu a exoforické odkazování jsou však zachyceny na tektogramatické rovině v jiných atributech než odkazování k jednoznačnému, explicitnímu antecedentu (viz III.7.1., III.7.2.), což nám umožňuje při popisu stávající anotace analyzovat tyto případy zcela samostatně, mimo kategorii textové koreference. Tím vymezíme textovou koreferenci jako vztah mezi dvěma koreferenčními výrazy a zachováme symetričnost a tranzitivitu vztahů.

Během anotace rozšířené textové koreference, anotace původní zájmenné koreference zůstala v zásadě stejná, až na některé drobné úpravy (k tomu viz III.8.).

III.4.2. Rozšířená textová koreference

Textová koreference anotovaná v současné době na PDT je rozšířením anotace pronominální textové koreference (III.4.1.). Toto rozšíření se týká především typů výrazu, na které se vztahuje anotace (koreferenční páry, kde anafor není vyjádřen osobním nebo ukazovacím zájmenem v substantivní funkci, ani není elidován (III.2.)). V následující kapitole se chceme věnovat principům, pravidlům a konvencím anotace rozšířené koreference (dále jen textová koreference) na tektogramatické rovině PDT. V oddíle III.4.2.1. je představena typologie textově koreferenčních vztahů s podrobným rozborem typů a ukázkami příkladů. Kapitola III.4.2.2. rozebírá textově koreferenční vztah z hlediska lexikálních skupin. V III.4.2.3. uvádíme a rozebíráme některé problematické případy označování koreference, přičemž zvláštní pozornost je věnována problematickým koreferenčním párům substantiv, která mají abstraktní význam (III.4.2.2.1.). Oddíl III.4.2.4. je věnován problematice správného určování antecedentů a obsahuje veškeré konvence a rozhodnutí výběru.

Při anotaci rozšířené textové koreference se nezaměřujeme na anaforické odkazování, ale pouze na identitu referentů antecedentu a anaforu (viz princip rozhodujícího koreferenčního vztahu v III.1.6.).

Aktuální informaci o anotovaných datech viz na webových stránkách projektu https://wiki.ufal.ms.mff.cuni.cz/anotace-rozsirene-koreference.

Textovou koreferenci anotujeme na vzdálenost nepřesahující 20 vět. Anotace koreference na větší vzdálenost v textu je přípustná pouze v případech automatické předanotace koreference pojmenovaných entit (III.4.2.2.3.). Při manuální anotaci na větší textovou vzdálenost neanotujeme především z důvodů předpokládaného velkého počtu chyb takové anotace. Anotátor si jen těžko vzpomene na antecedent, který se vyskytl v textu před více než 20 větami. Také z technického hlediska je náročné znázorňovat v anotačním editoru TrEd[92] více než 20 předchozích vět (textové okno zabere příliš mnoho místa, program je přetěžován informací atd.), tedy ve výsledku bude pravděpodobně více chyb, než správně označených souvislostí.

Z těchto důvodů označujeme koreferenci mezi NP *kanalizace* a *kanalizační sítě* v (3) a neoznačujeme ji mezi NP *situace* a *za krizovou situaci* v (4):

(3) *a. Poslanci budou muset odpočívat jinde , protože suterény jsou příliš hluboko a napojení na výše položenou kanalizaci pomocí čerpadel by provoz budov neúměrně prodražilo.*

[...17 vět...]

b. Existující kanalizační sítě {coref_text na „kanalizace"} *by totiž podzemní chodbu vtlačily tak hluboko do země, že by ji jistě nikdo nepoužíval.*

(4) *a. Situace začala být přirovnávána k porevolučním snahám některých západních firem „odložit" na naše území za malý peníz toxické odpady .*

[...44 věty...]

b. Její zástupce ing. Šedivý však veškerou odpovědnost za krizovou situaci {žádný koreferenční vztah} *odmítá.*

Textovou koreferenci **neoznačujeme** také v následujících případech:

1. Koreference otázkového slova a odpovědi v dialogických textech (*kde – zde, v Praze* aj., *kdy – dnes, v prosinci* aj.). Pro koherenci textu je vztah mezi otázkovým slovem a částí textu, která na tu otázku odpovídá, velice důležitý. Je to však jiný typ koheze textu, který již přesahuje tekrogramatickou rovinu a patří spíše do roviny diskurzu. Při rozšířené anotaci koreference na tektogramatické rovině tento vztah tedy nezachycujeme. Srov. např. (5)a–e, kde

[92] Anotační nástroj, ve kterém probíhají anotace na ÚFALu, viz Pajas – Štěpánek 2008 a IV.1.

souvislost mezi otázkou *kdy* a odpovědí mezi *16. až 18. hodinou*, *při změnách počasí*, *v obdobích před a po vysvědčení* a *v době viróz* se neoznačuje, i když tyto souvislosti výrazně podílí na kohezi textu:

(5) *a. Kdy děti nejvíce volají? [...]*
b. Podle zkušeností ze zahraničí se dá předpokládat, že největší frekvence telefonátů nastane vždy mezi 16. až 18. hodinou {žádný koreferenční vztah}.
c. A také při změnách počasí {žádný koreferenční vztah}, *které působí na citlivější organismus.*
d. V obdobích před a po vysvědčení {žádný koreferenční vztah}.
e. V době viróz {žádný koreferenční vztah}.

2. **Koreference zájmen první a druhé osoby** v dialogických a nedialogických textech. (k odůvodnění tohoto rozhodnutí viz III.2.1.1. bod 4.).

V dialogickém textu však běžně označujeme jiné vztahy, než koreference osobních zájmen 1. a 2. osoby a otázkové slovo – odpověď, jde-li o identickou koreferenci nebo asociační anaforu. Příklady jsou uvedeny v odpovídajících kapitolách bez zvláštního odkazu na to, že pochází z různých replik dialogického textu. Srov. např. (6)a–c:

(6) *a. Dovozoval, že vývoj kapitalismu se historicky vyznačuje dvěma fázemi: Fází soutěžního kapitalismu a fází kapitalismu trustů.*
b. Schumpeter se ve svém posledním díle ptá: „Který systém, kapitalismus {coref_text na „kapitalismus" v *a.*}, *či socialismus bude určovat budoucnost lidstva?"*
c. K údivu, úžasu či ohromení většiny svých kolegů odpovídá jednoznačně: „Bude to socialismus {coref_text na „socialismus" v *b.*}."

III.4.2.1. Typologie textově koreferenčních vztahů

Při anotaci rozšířené koreference na PDT 2.0 vycházíme z klasifikace typů reference podle Mendozové (2004) (viz přehled v II.2.). Rozlišujeme referenční a nereferenční jmenné fráze. Nereferenční jmenné fráze neanotujeme.

Za nereferenční tedy považujeme a neanotujeme:

- jmenné fráze v predikativní pozici, kromě případů identifikačních konstrukcí, kde jmenná část přísudku může sloužit antecedentem pro koreferenční odkaz v následujícím kontextu (viz III.1.4. a II.2.1.). Např. žádná koreference se neoznačuje mezi *Petr* a *programátor* v (7):

(7) *Petr je programátor.* (VL)

- jmenné fráze – druhé části apozičního spojení; tedy neanotujeme koreferenční vztah mezi členy apozice, např. mezi *Petr* a *programátor* v (8):

(8) *Petr, náš programátor, má zítra státnice.* (VL)

- výrazy, které mají řídící uzel s funktorem `ID`, protože jmenná fráze v takových případech nereferuje na objekt vnějšího světa, ale sama na sebe (tzv. autonymní použití v terminologii Padučevové (1985)).[93]
  Srov. nereferenční NP *Struktura* v (9)b:

(9) *a. Podle těchto zpráv nějaká firma na naše území umísťuje německou delikventní mládež, která zde páchá kriminální činy a ohrožuje starousedlíky.*
b. V Košťanech totiž zakoupila dům firma {coref_text na „firma" v *a.*} *Struktura* {funktor `ID`, žádný koreferenční vztah}*, která se u nás rozmísťováním německých chlapců zabývá.*

Srov. ale poněkud jiný příklad (10), kde výraz *převýchova* v (10)a je sice použit autonymně, ale vztah mezi NP *termín převýchova* v (10)a a *převýchova* v (10)b je anaforického typu a je relevantní pro koherenci textu. Proto ho označujeme jako nekoreferenční asociační anaforu (viz tento příklad ještě jednou v III.5.1.5.):

(10) *a. Pavel Vondráček: Termín* {`PAT`} *převýchova* {`ID`} *znám pouze z nacistického a komunistického slovníku.*
b. Na převýchovu {bridging na „termín" v *a.*} *se pokud vím, posílali ti, kteří měli podle těchto zrůdných režimů nevhodný původ.*

[93] Podobně se situace s identifikačními NP řeší v (Chiarcos – Krasavina 2005, s. 29). Avšak v jejich přístupu je řešení jednodušší, protože je prováděno na složkách, nikoliv na závislostní struktuře.

- jmenné fráze, které nemají v daném kontextu referenční platnost. Např. neanotujeme koreferenci u výrazů označujících měřítka, např. u uzlů jako `#Percnt`, *bod* apod. a kontextech jako (11):

(11) *a. Americký index obchodní důvěry odbytu a zaměstnanosti v příštích šesti měsících, se v srpnu snížil na 49,9 bodu, z 56,4 bodu* {žádný koreferenční vztah na „bod"} *v červnu.*
b. V dubnu byla jeho hodnota rovněž 49,9 bodu {žádný koreferenční vztah na „bod"}.

Ostatní použití jmenných frází považujeme za referenční s další diferenciací na předmětová jména se specifickou a nespecifickou referencí, abstraktní jména a dějová jména.

Textová koreference jako identita referentů v koreferefovaných párech výrazů není jev zcela jednotný z hlediska jeho identifikovatelnosti, resp. můžeme ho postulovat vždy s různou mírou přibližnosti pojmu koreference. Míra identifikovatelnosti koreferenčního vztahu záleží na typu reference a na sémantice daného výrazu:

1. Koreference jmenných frází se specifickou referencí a konkrétním významem je většinou zcela samozřejmá. Srov. např. koreferenci *maminka*$_a$ a *maminka*$_b$ v (12) – oba výrazy referují ke stejnému mimojazykovému objektu:

(12) *Helena poprosila maminku*$_a$, *aby na ni počkala. Maminka*$_b$ *však řekla, že nemůže. (VL)*

Stejně jasnou koreferenci zaznamenáváme u většiny konkrétních jmen se specifickou referencí, kde anafor nese sémantický rys určitosti.

2. Podobně vypadá koreference u **jmen referujících na konkrétní, ale nevybraný objekt**. Srov. koreferenci NP *kolega* a *ten* v (13):[94]

(13) *Poprosím o to některého kolegu a ten mi to řekne. (VL)*

[94] První polovina příkladu (do spojky *a*) je převzata z Adamce (1980). Adamec nazývá daný typ reference „podmíněně singulativní".

Stejné chování výrazů referujících na nevybraný objekt se specifickou referencí se vysvětluje tím, že jakmile podobná NP vstupuje do anaforického kontextu, dostává specifickou referenci (viz k tomu podrobněji v II.2.) a dále se chová jako běžná konkrétní jmenná fráze se specifickou referencí.

3. Jinak situace s koreferencí vypadá v případě, že **výrazy referují genericky**. Již samotný fakt generické reference není samozřejmý ani všemi lingvisty uznávaný (např. Berger (1993) mluví o generických NP jako o zvláštní skupině, stojící mezi referenčními a nereferenčními výrazy, Padučevová (1985) začleňuje generické NP mezi nereferenční, naopak Mendozová (2004) je pokládá za referenční, Helbig (2006) přisuzuje rys reference pouze negenerickým pojmenováním – viz v II.2.) atd. Reference na typ objektů se liší od reference na konkrétní vybrané objekty tím, že se nemusí vztahovat na všechny objekty daného typu. Srov. např. v (14) výraz *děti* nereferuje na všechny reálně existující děti, ale na prototypický pojem dítěte (vždyť jsou také děti, které nesnášejí čokoládu):

(14) *Děti milují čokoládu. (VL)*

Přirozeně vzniká otázka, zda můžeme považovat za koreferenční generické jmenné fráze, které odkazují na tentýž typ objektů. Následuje-li za (14) věta (15), výrazy *děti* v (14) a v (15) nemusí odkazovat na stejnou množinu dětí: jsou děti, které nesnášejí čokoládu, ale mají rády, když se jim čtou pohádky; nebo naopak některé děti nerady poslouchají pohádky, ale milují čokoládu.

(15) *Také mají děti rády, když jim rodiče čtou pohádky. (VL)*

Srov. také jiné možné pokračování věty (14):

(16) *Proto děti vždycky chtějí, aby jim ji maminka koupila. (VL)*

Ani v tomto příkladě, kde se v (16) mluví o dětech milujících čokoládu, množina objektů, na které referuje *děti* nemusí být totožná s množinou objektů, na které referuje *děti* v (14) – například ne všechny děti milující čokoládu vědí, co znamená koupit, ne všechny tyto děti mají maminku atd.

V našem projektu anotace rozšířené textové koreference **pokládáme za smysluplné anotovat textovou identickou koreferenci u generických jmenných frází,** a to z následujících důvodů:

1) Opakování stejného výrazu s generickou referencí se podílí na kohezi textu, podobně, jako je tomu v případě NP se specifickou referencí (srov. např. celý text (2) v IV.4. pojednávající obecně o dětech v dětských domovech, kde se v průběhu celého textu opakuje NP *děti* s generickou referencí);
2) Generické jmenné fráze se mohou zúčastnit veškerých anaforických vztahů a podobně jako jmenné fráze se specifickou referencí mohou být pronominalizovány (srov. paralelní syntaktické konstrukce v (17) a (18)), elidovány (19) nebo opakovány s ukazovacím zájmenem (srov. např. v (20) NP „tento podnikatel" s generickou referencí).

Pronominalizace:

- specifická reference NP *dítě*:

(17) *Moje dítě miluje čokoládu.* #PersPron *Vždycky chce, abych mu ji koupila. (VL)*

- generická reference NP *dítě:*

(18) *Děti milují čokoládu. Proto* #PersPron *vždycky chtějí, aby jim ji maminka koupila. (VL)*

- elipsa:

(19) *a. Tomu, kdo chce šetřit, hodně pomohou měřicí přístroje.*
b. #PersPron *Určí spotřebu a podle ní je zřejmě, co si lze dovolit.*

- opakování s ukazovacím zájmenem:

(20) *a. Tímto faktorem je podnikatel - inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk, ani ztrátu.*
b. Tento podnikatel {coref_text, na „podnikatel-inovátor"} *se od manažera liší tím, že zavádí nové kombinace výrobních faktorů, kdežto manažer je jen rutinně*

kombinuje na bázi dané techniky.

3) Hranice mezi NP se specifickou a nespecifickou referencí není vždy úplně zřetelná a v současné době není vůbec řešitelná automaticky. Pokud chceme naší anotací přispět k řešení aplikačních úkolů počítačové lingvistiky, měli bychom s tím počítat.

Uvedené argumenty nás vedou k označení vztahu mezi generickými jmennými frázemi odkazujícími na tentýž typ objektů jako koreferenční. Avšak toto rozhodnutí komplikuje skutečnost, že reference na typ objektů je mnohem složitější pojem, než specifická reference na konkrétní vybrané objekty. Jde o to, že typ není monolitním objektem a může mít potenciálně nekonečný počet podtypů, přičemž se na každý z podtypů může (i v rámci jednoho textu) referovat generický. Srov. např. řetězce generických NP v (21)–(22), jejichž referenty jsou postupně více specifikovány:

(21) *ženy – ženy v 19. století – české ženy v 19. století – bohaté české ženy v 19. století* atd.

(22)[95] *socialismus – socialismus v Německu – socialismus v Německu v 19. století*

Podobné případy se v textech vyskytují velice často, přičemž podtypy se mohou prolínat, křížit se mezí sebou a s celým typem. Většinou je poměrně složité a ani není vždy nutné je rozdělit a uspořádat. Jakmile nastává podobná situace, generické NP odkazující na různé podtypy, nebo na celý typ a jeho podtyp, už nejsou ani v uvedeném generickém smyslu koreferenční. Za koreferenční označujeme pouze takové páry generických jmenných frází, které odkazují na stejnou množinu objektů, čili snažíme se dodržovat extenzi koreferujících generických jmen. Vrátíme-li se k (21), koreferencí propojíme pár *ženy – ženy*, a *ženy v 19. století – ženy v 19. století*, nikoliv *ženy – ženy v 19. století.* Páry typu *ženy – ženy v 19. století* se dají v naší anotaci řešit pomocí asociační anafory typu „množina – podmnožina" (III.5.1.2.). V reálných textech se však setkáváme s případy, kdy toto rozlišení nelze provést důsledně. Tyto případy se řeší na základě intuice anotátora a jsou jednou z hlavních příčin nízké mezianotátorské shody v textech s velkým počtem NP s generickou referencí (viz příklady v IV.4.). O hraničních případech mezi generickými koreferenčními NP a generickými NP, které nemají být jako koreference zaznamenány viz III.4.2.3.2.

[95] Zde nejde v úzkém smyslu o generickou referenci, ale o odkazování na stejný příznak u abstrakt, s podobnými případy však zacházíme stejně jako s generickými NP.

Podíváme-li se na přístupy zpracování generické koreference v projektech z oblasti počítačové lingvistiky, vidíme, že např. Lezin (2007) při realizaci projektu automatického vyhledávání referenciálního propojení textu zahrnuje abstraktní NP spolu s generickými a predikativními NP do jedné skupiny „třídy objektů" a řeší je odděleně od jmenných frází se specifickou referencí. Navíc zvlášť vyčleňuje skupinu „třída objektů aktuální pro daný diskurz", kam spadají stále ještě generické NP, které jsou však o něco více specifikovány pro účely daného textu. Srov. např. (23)a a (24):

(23) *Žena nesmí do velké politiky.* (skupina „třída")

vs.

(24) *Žena v naší společnosti nesmí do velké politiky.* (skupina „třída aktuální pro daný diskurz")

Poesio (2000d) rozděluje generické a negenerické NP v definovaném pro jmenné fráze atributu GENERIC. Jako *generic-no* jsou označeny negenerické jmenné fráze, které referují ke konkrétním vybraným objektům, definovaným v určitém čase a místě. Příznak *generic-no* se přisuzuje především výrazům označujícím lidi, místa a konkrétní časové úseky (roky, staletí apod.). Jako *generic-no* se automaticky anotují rovněž jmenné fráze s identifikátorem určenosti *(the cat, this cat* apod.*)* a zájmena první a druhé osoby. Příznak *generic-yes* se přisuzuje výrazům odkazujícím k typům objektů (např. *tigers* v *Tigers are dangerous animals*). K *generic-yes* automaticky patří všechny NP s predikativním významem (v pozici přísudku nebo v apozici), ale také jmenné fráze v jiných syntaktických pozicích. Srov. např. (25)–(27):

(25) *I like music / wine / bread.*

(26) *The tiger / a tiger is a dangerous animal.*

(27) *The German / A German is a good musician.*

Většina výrazů s abstraktním významem se rovněž zařazuje do *generic-yes*, srov. *life* v *change of life*, *mythology* v *scenes from mythology* apod.

Třetí skupina *undersp-generic* se používá pro zaznamenání případů, kde generičnost jmenné fráze nelze spolehlivě určit, u koordinačních konstrukcí, kde se generické jmenné fráze koordinují s negenerickými a výrazů v modálních a neoznamovacích kontextech.

4. Další odlišný typ koreference je koreference abstrakt a dějových jmen. Podobně jako v případě s generickými výrazy, schopnost abstrakt a dějových jmen referovat je poměrně problematickou záležitostí (viz podrobněji v III.4.2.2.1 a III.4.2.2.2.). Pokud přesto uznáváme jejich schopnost referovat, zůstává otázkou, k čemu referují (srov. např. v Padučevové 1986) a jestli se mohou zúčastnit vztahů koreferenčních. Odpovíme-li na poslední otázku pozitivně, setkáme se s dalšími problémy: abstraktní jména jsou poměrně vágní a nepřesně vymezenou kategorií, která má složitou vnitřní hierarchii. Neexistují ani přesná kritéria pro odlišování abstraktních jmen od konkrétních (III.4.2.2.1.). V mnoha případech abstraktní jména, podobně jako jména dějová, mohou mít výrazný predikativní charakter, tedy nereferovat, ale obsahovat informaci o vlastnostech, proto například abstraktní a dějová jména nejsou lhostejná ke kategoriím času, místa apod. Podrobně se koreferenci abstraktních a dějových jmen věnujeme v III.4.2.2.1., tady chceme pouze upozornit na to, že informace o ontologickém statusu jména je pro analýzu koreferenčních párů velice důležitá.

Naše rozhodnutí ohledně (ne)anotace koreference u jmenných frází s různým ontologickým statusem a typem reference shrnujeme v tabulce č. 16.

referenční typ NP	**anotovat/neanotovat textovou identickou koreferenci**
specifická reference – konkréta	ANO
abstrakta	ANO
generické NP	ANO
predikativní NP	NE
nereferenční NP s funktorem ID	NE
nereferenční NP jako apozice	NE
jiná nereferenční NP (na nevybrány v diskurzu objekt)	ANO, pokud „koreferuje“ s anaforickou NP

Tabulka č. 16: Anotace textové koreference u NP s různou referenční platností

V ideálním případě potřebujeme mít pro anotaci koreference dodatečnou informaci o ontologickém statutu a typu reference. Takovou informaci však TGS neobsahuje a vytvořit ji nově nebylo prakticky možné. Proto v dané fázi anotace rozlišujeme dva druhy vztahů mezi koreferovaným antecedentem a anaforem v páru jmenných frází spojených textovou koreferencí – vztah mezi NP se specifickou referencí (typ=0) a vztah mezi NP s nespecifickou (především generickou) referencí (typ=NR)[96] (viz tabulku č. 17):

0	vztah mezi NP se specifickou referencí (viz B.2.2.1.)
NR	vztah mezi NP s generickou nebo nespecifickou referencí (viz B.2.2.4.)

Tabulka č. 17: Typologie textově koreferenčních vztahů

Při výběru mezi textovou koreferencí typu 0 a NR platí následující konvence:

Konvence o preferenci specifické reference u substantiv s primárně předmětným významem:

U koreferenčních jmenných frází s primárně předmětným významem v případě váhání mezi specifickou (typ=0) a generickou (typ=NR) referencí, preferujeme „defaultní" typ 0.

Na začátku anotace jsme rozlišovali také typy SYN (od synonymum) pro vztah mezi koreferenčními jmennými frází se specifickou referencí, které jsou vyjádřeny různými řídícími lexémy) a ER (od hyperonymum) pro vztah mezi NP se specifickou referencí, kde anafor je lexikální hyperonym ve vztahu k antecedentu. Tímto způsobem je oanotováno cca. 10 procent PDT. Potom se však ukázalo, že vztah hyponym/hyperonym je velice nejednoznačný, dostává se tam mnoho sporných případů, které zhoršují mezianotátorskou shodu a prodlužují dobu anotování; ve své čisté podobě se hyperonmie mezi koreferenčními NP se specifickou referencí vyskytuje v anotovaných textech jenom v jednotlivých případech. Rozlišování mezi tzv. přímou anaforou (opakující se NP se stejným řídícím členem) a koreferenčními páry, kde řídící uzel anaforu je vyjádřen jiným lexémem než řídící uzel antecedentu (typ SYN), se dá provést automaticky s použitím atributů tektogramatické roviny.

[96] NR – od nespecifická reference.

III.4.2.1.1. Koreferenční vztah mezi výrazy se specifickou referencí (coref_text, typ=0)

Prototypický koreferenční vztah daného typu je koreference dvou jmenných frází se specifickou referencí (odkazování ke konkrétnímu existujícímu, reálnému referentu a objektu skutečnosti.). Srov. koreferenci výrazů *smlouva* v (28)a–b:

(28) *a. V praxi to znamená, že i kdyby hnedka zítra řekla ČR, že smlouva je pasé, přesto by se teprve v březnu příštího roku mohla legislativně zbavit svých závazků vůči partnerovi z bývalé ČSFR.*

b. Na pozadí vývoje v posledních dnech a týdnech se však zdá, že litera výše uvedené mezinárodní smlouvy {coref_text, typ=0 na „smlouva" v *a.*} *mezi ČR a SR bude mít co nevidět pouze sílu psaného slova a ničeho jiného.*

Koreferenčního vztahu se specifickou referencí (typ 0) se mohou zúčastnit následující páry výrazů:

a) Původní pronominální koreference

Všechny vztahy původní pronominální koreference mají předvolený typ 0. V případech, kdy to neodpovídá skutečnosti, typ vztahu se následně ručně opravuje (viz III.8.).

- Antecedentem je zájmeno nebo rekonstruovaný uzel (s t-lemmaty `#PersPron`, `#Cor`, `#QCor` aj.)
- Rozšířenou koreferencí typu 0 se nejčastěji „dotváří" koreferenční řetězce původní pronominální koreference, tj. spojování párů typu `#PersPron` – *NP* při existujících párech typu *NP* – `#PersPron` (viz příklady a vysvětlení v III.2.). Srov např. doplňování vztahu *`#PersPron` – Péťa* při již existující pronominální koreferenci *Péťa* – `#PersPron` v (29)a–c.

(29) *a. Sedmiletý Péťa se půl roku neuvěřitelně trápil, že má AIDS.*

[...4 věty...]

b. `#PersPron` {coref_text, typ=0 na „Péťa" v *a.*} *Stále na to myslel, ve škole se už nedokázal soustředit.*

[...5 vět...]

c. Péťa {coref_text, typ=0 na `#PersPron` v *b.*} *skončil u Jany Drtilov.*

Srov. také propojování rozšířené textové koreference *Křesťanská misijní společnost – Společnost* s gramatickou koreferencí *Křesťanská misijní společnost – která* v (30):

(30) *a. Informovala o tom Křesťanská misijní společnost, která* {coref_gram na „společnost"} *toto shromáždění pořádala.*
b. Společnost {coref_text, typ=0 na „který" v *a.*} *vznikla v roce 1989 jako platforma pro spolupráci různých křesťanských směrů.*

b) **Opakování stejného pojmenování**

Anafor je formálně identický uzel s antecedentem. Srov. koreferenci NP *soutěž* v (31)a–b:

(31) *a. Jeho dojetí znásobila při vyhlašování přítomnost [...] pořadatelů soutěže – Českého manažerského centra v Čelákovicích.*
b. Na letošním ročníku soutěže {coref_text, typ=0 na „soutěž" v *a.*} *se spolupodílí i Profit .*

c) Opakování stejného pojmenování s determinátorem

Vyjádření identity pomocí textových identifikátorů. Příklad, kdy se se stejnou NP v anaforické pozici používá ukazovací zájmeno. Srov. (32):

(32) *Ten článek v dnešních novinách o otci, který utekl od ženy a dětí, aby je nemusil živit, to je strašné. Co bude teď chudák ta žena* {coref_text, typ=0 na „žena"} *s dětmi dělat?*

d) **Opakování různých podstromů při stejném řídicím uzlu**

Jako coref_text, typ=0 označujeme také případy, kde opakování antecedentní NP je částečné. Např. řetězce *společnost – akciová společnost – společnost Incheba*; *Vlček – ředitel J. Vlček – Jiří Vlček; ministr financí – ministr – tento ministr* atd.

Srov. např. *Ministerstvo financí – Ministerstvo financí ČR* v (33):

(33) *a. Nejvíce Ministerstvo financí .*
b. Nejvíce se na tom podílel resort Ministerstva financí ČR {coref_text, typ=0 na „Ministerstvo financí" v *a.*} *– a to formou daňových úlev ve výši zhruba 7,5 miliardy korun.*

e) **Antecedent a anafor jsou různá pojmenování**

Antecedent a anafor řídících uzlů koreferenčních podstromů jsou lexikálně vyjádřené autosémantické jmenné fráze s různými t-lemmaty. Existují následující možnosti:

- antecedent a anafor jsou synonymické:

(34) *a. Chlap je z Prahy, klidně může zasedat v koordinačním centru nebo být poradcem bůhví koho, takže pozor, tím vtipkováním si tě taky může prověřovat...*
b. Skřípavě se zasmál a řekl: A taky, chválabohu, hned tak nepochováme.
c. Ten hoch {coref_text, typ=0 na „chlap" v *a.*} *má tuhý kořínek, ten má sílu, ten má elán...* (Frýbová, Z., Hrůzy lásky a nenávisti)

- antecedent a anafor jsou jiná pojmenování než synonymická v úzkém smyslu. Srov. v (35)a–b výrazy *materiál* a *dokument* nejsou synonymické v úzkém smyslu, ale referují k témuž mimojazykovému objektu:

(35) *a. Jak je dále v materiálu zdůrazněno, pozitivní posun v rozvoji malých a středních podniků byl umožněn především díky stabilnímu makroekonomickému prostředí, relativní legislativní stabilitě a státní politice podpory podnikatelských subjektů.*
b. Z dokumentu {coref_text, typ=0 na „materiál"} *dále vyplývá, že v roce 1993 bylo celkově na podporu zejména malého a středního podnikání poskytnuto z rozpočtových prostředků více než 11 miliard korun.*

- anafor je v hyperonymickém vztahu k antecedentu, tj. pro anaforickou jmennou frázi se vybírá obecnější substantivum, než to, které je použito pro pojmenování antecedentu. Anaforická jmenná fráze se v takových případech používá většinou s ukazovacím zájmenem. Srov. koreferenci mezi *ÚNMS* a *tento úřad* v (36)a–b:

(36) *a. Usnesením vlády SR je koordinací všech akcí souvisejících se zajištěním certifikace dovážených potravinářských výrobků pověřen ÚNMS SR.*
b. Na tomto úřadě {coref_text, typ=0 na „ÚNMS SR"} *lze získat i potřebné informace.*

- antecedent a anafor jsou v relaci obecné jméno – pojmenovaná entita:

(37) *a. V praxi to znamená, že i kdyby hnedka zítra řekla ČR, že smlouva je pasé, přesto by se teprve v březnu příštího roku mohla legislativně zbavit svých závazků vůči partnerovi z bývalé ČSFR.*

b. Na pozadí vývoje v posledních dnech a týdnech se však zdá, že litera výše uvedené mezinárodní smlouvy mezi ČR a SR {coref_text, typ=0 na „partner"} *bude mít co nevidět pouze sílu psaného slova a ničeho jiného.*

f) Antecedentem koreferenčního vztahu je sloveso, propozice nebo věta

Jako antecedent může vystupovat slovesná fráze, propozice, celá věta nebo dokonce několik vět. V případě odkazu k několika větám použijeme speciální typ odkazování k segmentu textu (atribut coref_special, typ=`segm`; podrobněji viz III.7.2.). V ostatních případech šipka vede na řídící výraz antecedentu, který zastupuje celou větu. Odkazování daného typu označujeme jako textovou koreferenci, typ 0 (`coref_text, typ=0`). Srov. (38):

(38) *a. Podle regulí GATT lze toto opatření přijmout maximálně na období šesti měsíců a pouze u vybraných položek.*

b. Tato skutečnost {coref_text, typ=0 na řídící výraz antecedentní věty „lze"} *však nic nemění na faktu, že nadcházející týdny a měsíce budou znamenat neúměrně zvýšené nároky na administrativu podnikatelů při rozvíjení jejich obchodních aktivit se slovenskými partnery.*

g) Nespecifická negenerická koreference

Jako koreferenci typu 0 anotujeme také případy párů jmenných skupin s nespecifickou ale přitom negenerickou referencí v případě, že anafor je použit s určitým identifikátorem. Jde o takový typ reference, kdy objekt sice referuje ke konkrétnímu objektu dané třídy, ale tento objekt ze třídy není vybrán (podle klasifikace Padučevové má nereferenční existenciální denotační status, viz II.2.). V daném případě antecedent má sice nespecifickou referenci, ale pak se s ním operuje jako s konkrétním vybraným objektem, čili vytváří se fiktivní svět daného diskurzu, ve kterém se daný objekt chová jako existující a reálný. Srov. např. koreferenci výrazů *podnik* a *úřad* v (39), a *velice pozorný člověk* v (40):

(39) *Například muž, který pracuje v nějakém velkém podniku, se zakouká do sekretářky ve stejném podniku* {coref_text, typ=0 „podnik"} *a začnou se scházet v nějaké*

kavárničcce stranou od toho úřadu {coref_text, typ=0 „podnik"}.

(40) *Přesto si značky mohl všimnout jen někdo velice pozorný [...] a ani ten velice pozorný člověk* {coref_text, typ=0 „někdo"} *by jim patrně nepřikládal žádný význam.*

Kataforický odkaz dopředu

Kataforický odkaz dopředu anotujeme pouze v případě skutečné textové katafory (ve smyslu Berger 1993; Mendozová 2004, s. 118). Srov. např. (41)–(42):

(41) *a. Tu nejvhodnější dobu* {coref_text, typ=0 na „rok" v *b.*} *pan Hrabák propásl.*
b. V osmdesátých letech se daly pořídit krásné věci za, viděno dneškem, ještě krásnější ceny.

(42) *a. Na převýchovu se pokud vím, posílali ti, kteří* {coref_text, typ=NR na #Comma v *b.*} *měli podle těchto zrůdných režimů nevhodný původ.*
b. Židé, cikáni, šlechta, podnikatelé, kulaci a jiní.

III.4.2.1.2. Koreference generických jmenných frází (coref_text, typ=NR)

O jmenné skupině se říká, že je použita genericky, jestliže jejím referentem je klasický, vzorový, prototypický představitel dané třídy (např. (43)–(45)), odkaz na libovolný element dané třídy (např. (46)) nebo na representativní podmnožinu referentů třídy (např. (47)) (viz II.2.).

(43) *Honza dokáže zabít vlka.*

(44) *Mokrá veverka vypadá jako myš.*

(45) *Kočka má zelené oči.*

(46) *Členové tohoto klubu nepijí whisky.*

(47) *Američané přistáli na Měsíci v r. 1969.*

Avšak vymezení generických NP zdaleka není jednoduchou záležitostí a ve skutečných textech se často vyskytují problematické případy. K vymezení generických NP používáme dva praktické heuristické testy:

- test na generickou referenci Rachilinové – Krejdlina (1981):
  jmenná fráze *X* je generická, pokud *X* může být použit v konstrukci „*X* jako <typický> *Y*", „*X* jako druh (forma) *Y*". Například ve větě (48) *vlak* má generickou referenci, protože (48) můžeme přeformulovat na (49):

(48) *Jezdí vlakem.*

(49) *Jezdí vlakem, protože je to nejlevnější dopravní prostředek.*

Tento test však nemůže být použit ve většině případů, kdy testovaná jmenná fráze je v textu v plurálu, např. těžko vymyslíme podobnou formulaci pro (50):

(50) *Děti milují hračky.*

- naše praktická anotátorská pomůcka:
  jmenná fráze X je generická, pokud ji můžeme převést do kontextu, kde bude použita predikativně. Například ve větě (51) NP *děti* je použita genericky, protože ji můžeme přeformulovat jako (52):

(51) *Děti mají rady zmrzlinu.*

(52) *Pokud X je dítě, x má rád zmrzlinu. nebo Ti, kdo jsou děti, mají rady zmrzlinu.*

Argumenty pro zaznamenávání koreference u generických NP se stejnou extenzí uvedené v III.4.2.1. vedou k formulaci následujícího pravidla:

Pravidlo o anotaci textové koreference u generických NP:

Textovou koreferenci typu NR anotujeme u jmenných frází s generickou referencí, pokud referují ke stejnému typu objektů stejného rozsahu.

Textovou koreferenci typu NR zaznamenáváme:

1. U generických jmenných frází v singuláru a v plurálu

a) **vyjádřených stejným pojmenováním**, srov. NP *českým exportérům* v (53):

(53) *a. Nová striktní omezení vlády SR proti českým exportérům.*
b. Již několik dnů je všeobecně známo, že ochranářská opatření slovenské vlády proti českým exportérům {coref_text, typ=NR na „exportér" v (53)a} *se dotýkají zejména oblasti obchodu s potravinami a zemědělskými produkty.*

b) pokud anaforický člen je **pronominalizace nebo aktuální elipsa generického antecedentu**. V tomto případě měníme automaticky předvolený typ 0 v anotaci původní pronominální anotaci na typ NR.[97] Srov. aktuální elipsu NP *měřící přístroje* v (54) a pronominalizaci NP *droga* v (55):

(54) *a. Tomu, kdo chce šetřit, hodně pomohou měřicí přístroje.*
b. #PersPron {coref_text, typ=NR na „přístroje" v (54)a} *Určí spotřebu a podle ní je zřejmě, co si lze dovolit.*

(55) *Droga je tedy tak účinná, že ten, kdo ji* {coref_text, typ=NR na „droga"} *užívá, se snadno dostane do „pohody" kouřením nebo šňupáním.*

c) **antecedent a anafor jsou různá** (např. synonymní) **pojmenování**. Srov. např.:

(56) *a. Na telefonní číslo 855 44 33 bude jistě volat mládež s různými problémy .*
b. Doufejme, že linka si časem vydobude mezi dětmi {coref_text, typ=NR na „mládež" v (56)a} *takovou autoritu, aby se na ni obracely i ty, které jsou skutečně ohrožovány.*

d) kde **jeden z členů páru je zkratka** a jiný je rozepsaná zkratka. Srov. např. (57):

(57) *a. O odpočtu DPH.*
b. Podle novely zákona o dani z přidané hodnoty {coref_text, typ=NR na „DPH"} *se letos stanu plátcem daně.*

[97] Toto by bylo možné provést i automaticky, avšak ztratila by se informace o prolínajících se koreferenčních řetězcích se specifickou a generickou referencí (viz IV.1.2.)

e) pokud **anaforická generická NP je hyperonym ve vztahu k antecedentu**. V tom případě platí pravidlo, že použití aktualizátoru pro zachování koreference s antecedentem je nezbytné. Srov. např. (58):[98]

(58) *S příchodem jara sníh odtál a Vítězslav mohl nechat dřevo konečně odvézt, než do něj nalétne kůrovec. Byl začátek května, teplého května a již se ten malý brouček, ale velký škůdce lesa* {coref_text, typ=NR na „kůrovec"}, *začínal rojit.*

f) **u uzlů závislých na „kontejnerech",**[99] resp. u uzlů s funktorem MAT, které považujeme za generické (*sklenice mléka* apod.) (viz III.4.2.3.3.). Srov. koreferenci generických NP *surovina* a *heroin* v (59)a–b a na obrázku č. 14:

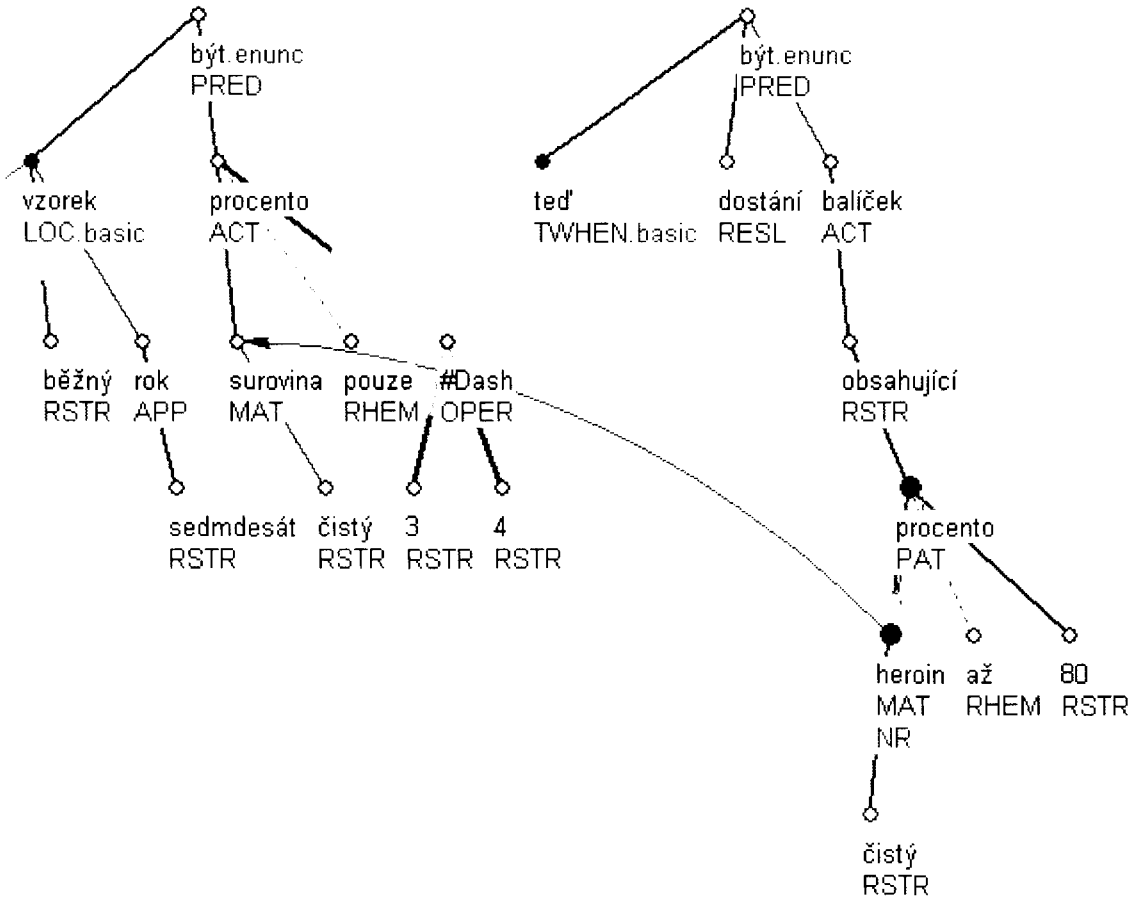
(59) *a. V běžném vzorku sedmdesátých let byla pouze 3–4 procenta čisté suroviny.*
b. Nyní jsou k dostání balíčky obsahující až 80 procent čistého heroinu {coref_text, typ=NR na „surovina"}.

Srov. také příklad (60), kde oba koreferenční generické uzly jsou závislé na kontejnerech se specifickou referencí:

(60) *a. Křesťané se modlili za usmíření národů...*
b. Více než tisícový zástup křesťanů {coref_text, typ=NR na „křesťan" v (60)a} *z různých sborů a církví českých zemí a delegace křesťanů* {coref_text, typ=NR na „křesťan" v (60)b, funktor APP} *z Německa se v sobotu na vrchu Radobýl u Litoměřic modlil za smíření mezi Čechy a sudetskými Němci.*

[98] Příklad ze SYN2000.

[99] K výsvětlení pojmu "kontejner" v anotaci tektogramatické roviny viz Mikulová a kol., 2005:809.



Obrázek č. 14: Koreference generických výrazů závislých na kontejnerech

2. U většiny jmenných frází s abstraktním významem (podrobný rozbor anotace koreference u jmenných frází s abstraktním významem viz v III.4.2.2.1.). Srov. např. (61):

(61) *Tímto faktorem je podnikatel – inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk* {coref_text, typ=NR na „zisk"}, *ani ztrátu.*

3. U negenerických nereferenčních jmenných frází, pokud se v pozici antecedentu a anaforu objeví tatáž nereferenční jmenná fráze se stejným typem nespecifické reference a se stejnou extenzí, ale bez determinátoru (v případě anaforického opakování nereferenční NP viz III.4.2.1.1.). Nebude to ovšem anaforický vztah, ale pouhá koreference. Srov. (62)a–b:

(62) *a. Když si dítě bude přát, aby se o jeho problému nikdo z rodiny nebo školy nedozvěděl, musíme to respektovat, vysvětluje Jana Drtilová . [...]*

b. Většinou se stává, že dítě ani nechce, aby se rodina {coref_text, typ=NR na „rodina" v a.} *dozvěděla, že se nám ozval.* (VL)

c. Linka by neměla rodinu {žádný koreferenční vztah} *nahrazovat, ale doplňovat.*

Pokud však věta (62)a pokračuje větou (62)c, koreferenční vztah mezi *rodina* v (62)c a *rodina* v (62)a neanotujeme. V (62)a *rodina* má nereferenční negenerickou interpretaci, zatímco v (62)c *rodina* je použita genericky. Ani při poměrně širokém chápání pojmu koreference tyto dvě NP nejsou koreferenční.

Je zřejmé, že vztah mezi negenerickými nereferenčními NP nepřispívá tolik koherenci textu a možná by bylo logičtější takové vztahy vůbec nezaznamenávat – koreference v úzkém smyslu zde není (obě NP referují na nevybraný objekt), o anaforický vztah také nejde (jinak by se to označovalo jako koreference NP se specifickou referencí, viz III.4.2.1.1.). Avšak hranice mezi nereferenčními negenerickýcmi a generickými jmennými frázemi v neanaforickém kontextu je velice vágní, a provádět tuto hranici uměle na základě konvencí by byl další časově náročný a v podstatě zbytečný úkol. Proto můžeme daný vztah označovat jako koreferenci s typem NR.

III.4.2.1.3. Koreferenční řetězce s prolínající se specifickou a nespecifickou referencí

V textu se koreferenční řetězce typu 0 a typu NR mohou prolínat, tj. některé hrany jednoho řetězce mohou být označeny NR, jiné však 0, zejména tehdy, kdy se v řetězci střídají různá pojmenování. Srov. příklad dlouhého hypertematického řetězce v (63)a–e:

(63) *a. Také lidé z okolních domů si stěžovali na hluk, výtržnosti, aroganci a proudy holek, které se za kluky táhly.*

b. Bylo jim však divné, že chlapce {specifická reference, coref_text, typ=0 na „kluk" v *a.*} *nikdo nevede, nehlídá ani nevychovává.*

c. Kdo to {specifická reference, coref_text, typ=0 na „chlapec" v *b*} *vlastně je?*

d. Německé chlapce {nespecifická reference, coref_text, typ=NR na „ten" v *c.*} *jsme již nezastihli .*

e. Duchcov byl posledním místem, odkud #PersPron {specifická reference,

coref_text, typ=0 na „chlapec“ v *d.*} *byli těsně před naším příjezdem odvezeni zpět do Německa.*

V následujícím případě (64)a–d NP *chicle* má třikrát generickou (64)a–c a jednou specifickou (64)d referenci.

(64) *a. Tak jako každý Mexičan , i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle, a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku.*
b. Santa Anna má chicle {coref_text, typ=NR na poslední „chicle“ v *a.*} *a Adams technické schopnosti.*
c. Asi rok se Adams a jeho nejstarší syn snažili – chicle {coref_text, typ=NR na „chicle“ v *b.*} *vařili, čistili, přidávali množství různých látek a míchali s pravým kaučukem.*
d. Když asi po roce své úsilí vzdali, rozhodl se Adams, že vše, co mu z chicle {coref_text, typ=0 na „chicle“ v *c.*} *ještě zbylo, hodí do řeky.*

Srov. také koreferenci v páru *Romové – tento národ*, kde první výraz má generickou referenci a druhý – specifickou:[100]

(65) *a. Nic z toho se však nevyrovná míře neštěstí, které Romy* {nespecifická reference} *postihlo v letech druhé světové války.*
b. Spolu se Židy #PersPron {nespecifická reference, coref_text, typ=NR na „Romy“ v *a.*} *byli označeni za méněcennou rasu a stali se objektem patologických fašistických opatření, jejichž cílem byla úplná genocida tohoto národa* {specifická reference, coref_text, typ=0 na #PersPron v *a.*}

III.4.2.2. Textová koreference z hlediska lexikálních skupin

III.4.2.2.1. Koreference abstraktních jmen

Jedním z nejproblematičtějších bodů v anotaci rozšířené textové koreference je zpracování abstraktních jmen. Substantiva s abstraktním významem stojí na pomezí mezi referujícími jmény s předmětovým významem a predikujícími slovními druhy, jako jsou např. adjektiva, adverbia a slovesa. Avšak zatímco u jiných slovních druhů bylo možné stanovit alespoň relativně formální kritéria pro výrazy podléhající anotaci (viz III.2.), v případě abstraktních substantiv to zdaleka není tak jednoduché.

[100] Vysvětlení, proč NP *tento národ* považujeme za specifickou viz v III.4.2.3.1.

Základní, jednoduchá definice je, že konkrétní substantiva jsou ta, která označují hmotné věci, např. *strom, kámen, papír, vlasy...* . Naopak abstraktní substantiva mají význam nehmatatelných objektů, např. *pocit, strach, láska, představivost...* .

Rozdělení lexika na abstraktní a konkrétní je zásadní (srov. už Frege 1892). Obě třídy (abstraktní a konkrétní) jsou však poměrně dynamické a není vyloučeno, že u některých jmen nebude zcela zřejmé, kam je zařadit.

Klasifikace lexika na abstraktní a konkrétní a pohled na referenční vlastnosti abstraktních jmen jsou u různých autorů velice odlišné.

Ju. S. Stepanov dělí jména na denotátní a signifikátní. Denotátní slovní zásoba se směruje k označení reálných předmětů vnějšího světa, denotátů, zatímco signifikátní slovní zásoba spíše pojmenovává pojmy, signifikáty (Stepanov 2004, s. 59). K denotátním patří také obecné termíny, které se determinují výčtem součástí podle principu „část – celek". Obecný termín je názvem určité situace, závislé termíny vytváří tematickou třídu, jako např. *oblečení* (obecný termín) – *sukně, košile, ponožky* apod. (závislé termíny). Signifikátní jména jsou např. *zvíře* jako obecný název pro množinu *vlk, kráva, kůň* apod., mají strukturní vztahy třída – jednotka, elementy této třídy mohou podle Štěpanova vždy zaměnit svůj hyperonymum. Jména obou zmíněných tříd mohou mít ve výpovědi konkrétní denotát, signifikátní jména se však používají i v kontextech, kde je možné je interpretovat jako abstraktní.

Velice střízlivý a prakticky aplikovatelný přístup nacházíme u Ufimcevové (1986). Podle ní je rozdělení na konkrétní a abstraktní lexiku graduální a řeší se podle toho, který komponent významu – denotátní nebo signifikátní – u daného slova převládá. Pokud převládá denotátní aspekt, jde o konkrétní jméno, pokud ve významu slova převládá nebo je přítomen pouze signifikátní aspekt, jde o abstraktní jméno. Při takové klasifikaci se za konkrétní považují počitatelné předměty, osoby, zvířata, ale také nepočitatelné hmoty.

Aruťunovová (1976) vychází ze syntaktických vlastností jmen, především z míry jejich syntaktické svobody, a podle toho klasifikuje jména na tři skupiny: osoby, neosobní konkrétní předmětová jména a abstraktně-událostní jména, tj. osoby jsou chápány na škále abstraktní – konkrétní jako nejvíce konkrétní.

Černějková (1997) tvrdí, že rozlišení abstraktního a konkrétního lexika není jednoznačné a má spíše graduální povahu. Velmi obecně jako abstraktní rozumí výrazy myšlenkové, pojmové, zatímco konkrétní mají předmětný věcný obsah. Černějková nabízí několik principů, podle kterých toto rozdělení většinou provádíme:

1. Referenční princip: konkrétní jména odkazují na věci hmotné podstaty, abstraktní jména takový denotát nemají.
2. Formálně sémantický princip: abstraktní jsou taková jména, která popisují vlastnosti, stavy a vztahy mezi věcmi zvlášť od jejich hmotných nositelů.
3. Sémantický princip: abstraktní jsou jména, která mají širší význam ve srovnání s jinými slovy, a která jsou propojená s těmito slovy vztahem třída – představitel.
4. Syntaktický princip: abstraktní slova jsou častější v predikativní funkci.

Celkově je možné říci, že všichni uvedení autoři věnující se problematice konkrétních a abstraktních jmen hodnotí hranici mezi nimi jako graduální a hledají systém kritérií, jehož použití pomáhá zařadit jednotlivé výrazy do jedné z těchto kategorií.

Použijeme-li představené názory, budeme schopni ve většině případů identifikovat abstraktní jména. Čeká nás však další zásadní problém – posoudit jejich referenční schopnosti a podle toho se rozhodnout, ve kterých případech podléhají anotaci koreference.

E. V. Padučevová věnuje několik článků referenčním schopnostem deverbativních (dějových) substantiv a jiných jmen s nepředmětovým významem.[101] Její základní myšlenka spočívá v tom, že na rozdíl od predikátů, které nemají vlastní referenci, celé propozice (včetně svých aktantů), a to na situaci, kterou pojmenovávají. Typ reference propozičních komponentů se definuje na základě několika parametrů: odkaz na situaci nebo na fakt (možnost), reálná a neutrální modalita apod. Je to však zcela jiný přístup k referenci, než v případě jmen s konkrétním významem.

I. Mendozová (2004, s. 160n) v referenci abstrakt nevidí zásadní rozdíl od referencí předmětových jmen. Stejně jako předmětová substantiva, abstrakta mohou mít specifickou i generickou referenci. Rozdíl je však v tom, že podle Mendozové abstrakta jsou valenční, tedy pokud valence abstraktního substantiva má specifickou referenci, potom ji má i řídící abstraktní substantivum (srov. *tvrdohlavost* aplikovatelná ke konkrétnímu nositeli této vlastnosti v (66)), v opačném případě výraz má generickou referenci (67)).

(66) něm. *Ihr Starrsinn brachte ihn zur Verzweiflung.*
č. *Její tvrdohlavost ho přivedla k zoufalství.*

(67) něm. *Liebe ist ein Gefühl.*

[101] Viz Padučevová 1979, 1983, 1986.

č. *Láska je pocit.*

Podobně jako pro jiná substantiva, pro abstrakta (a deverbativa) je relevantní opozice určitosti – neurčitosti. Avšak u abstrakt splývá určitost (*x – ten x*) a tzv. kvalitativní neurčitost (*x – takový x*), srov. příklad z Padučevové (1988):

(68) rus. *Маша покраснела, отвернула немного голову в сторону и в таком / этом положении продолжала говорить.*
č. *Maša zčervenala, pootočila trochu hlavu do strany a v této / takovéto poloze mluvila dál.*

Srov. také příklad z PDT s možností záměny *tento* na *takový* v (69)b:

(69) *a. Slovenská celní správa bude vyžadovat u každé jednotlivé dodávky originální rozhodnutí či certifikát.*
b. Tento postup (= takový postup) si vyžádá v praxi zhotovování ověřených kopií.

* * *

Naše řešení pro anotaci koreference u substantiv s abstraktním významem kombinuje snahu o zachování důslednosti v anotaci specifické a generické koreference a konvenci pro dosažení co nejvyšší mezianotátorské shody. U abstrakt se totiž snažíme rozlišovat specifickou a nespecifickou referenci, ale děláme to takovým způsobem, že pokud daná NP má jasně specifickou referenci, přiřadíme jí typ=0. Pokud však rozhodování o typu reference dané jmenné fráze není jednoznačné a vyžaduje aspoň trochu přemýšlení, označíme ji jako coref_text, typ=NR. Uvedeme několik příkladů.

V (70) se může zdát poněkud zvláštní propojovat *strach* v (70)a a (70)b textovou koreferencí, protože obě jmenné fráze jsou v rématu a nejde o anaforické odkazování. Antecedent v (70)a má ještě navíc funktor CPHR, což také „ochuzuje" jeho referenční status. Avšak vzhledem k tomu, že jde o tentýž děj (skutečnost, které se dítě bálo), spojíme ho zde koreferencí.

(70) *a. Přiznal, z čeho má strach. [...]*

b. Všechno nakonec dobře dopadlo, ale tohle dítě zbytečně prožilo půl roku strachu a děsivých představ {coref_text, typ=0 na „strach" v *a.*}.

Srov. také (71)a–b pro specifickou koreferenci abstraktního jména *ekonomika* s konkrétním aktantem *Česko*:

(71) *a. Ve specifických podmínkách české ekonomiky růst nezaměstnanosti v letech 1991 – 1993 značně zaostal za poklesem HDP.*

[...7 vět...]

b. Nejméně dvouprocentní růst české ekonomiky {coref_text, typ=0 na „ekonomika" v *a.*} *již letos (1999).*

Srov. dále příklady (72)a–b s generickou referencí:

(72) *a. Dovozoval, že vývoj kapitalismu se historicky vyznačuje dvěma fázemi: Fází soutěžního kapitalismu a fází kapitalismu trustů.*

b. Schumpeter se ve svém posledním díle ptá: Který systém, kapitalismus {coref_text, typ=NR na „kapitalismus"}*, či socialismus, bude určovat budoucnost lidstva?*

c. K údivu, úžasu či ohromení většiny svých kolegů odpovídá jednoznačně: Bude to socialismus {coref_text, typ=NR na „socialismus"}.

Páry *kapitalismus – kapitalismus* a *socialismus – socialismus* zařazujeme do generické koreference typu NR z několika důvodů:

- můžeme na ně použít náš test pro určování generické reference (viz v III.4.2.1.),
- nemá rozvití se specifickou referencí (podle Mendozové 2004),
- při anotaci nemůžeme rozhodnout, jak s takovým párem naložit, což je také kritérium pro typ NR.

V (73) je situace odlišná. Také tady jsme se rozhodli pro generickou referenci, avšak z poněkud jiných důvodů. Výraz *zisk* má predikátovou povahu (má alespoň dvě intenční místa) a může mít jak abstraktní tak konkrétní interpretaci. V případě abstraktní interpretace je citlivé na kategorie času a dosahu. V daném případě jde zřejmě o interpretaci abstraktní a generickou.

(73) *a. Tímto faktorem je podnikatel – inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk* {coref_text, typ=NR na „zisk"}, *ani ztrátu.*

b. Na konci tohoto difusního procesu se systém vrátí ke statické rovnováze, v níž nebudou opět ani zisky {coref_text, typ=NR na „zisk" v *a.*}, *ani ztráty.*

Následující příklad (74)a–c je velmi typický pro texty PDT, dokonce i s výrazem *nezaměstnanost.* V takových případech považujeme jmenné fráze s abstraktním významem závislé na výrazech typu *míra, růst* apod. za generické a anotujeme jejich koreferenci s podobnými NP, i v případě, že řídící substantiva jim přidávají různé časové a místní charakteristiky. Srov. v řetězci (75) v (74)a–c anotujeme generickou koreferenci pro *nezaměstnanost* ačkoliv nezaměstnanost je pokaždé jindy a jinde.

(74) *a. Míra nezaměstnanosti by se měla vyvíjet protikladně, než ve standardní ekonomice.*

b. Ve specifických podmínkách české ekonomiky, mj. vzhledem k netržnímu chování neprivatizovaných podniků, nízkým mzdám jakož i rychlému rozvoji drobné podnikatelské aktivity, růst nezaměstnanosti {coref_text, typ=NR na „nezaměstnanost" v *a.*} *v letech 1991 – 1993 značně zaostal za poklesem HDP.*

c. Pokračující privatizace a restrukturalizace si však vynutí zvýšení míry nezaměstnanosti {coref_text, typ=NR na „nezaměstnanost" v *c.*} *z 3,5% koncem roku 1993 na 5 – 6 % ke konci příštího roku.*

(75) *míra nezaměstnanosti – růst nezaměstnanosti v letech 1991 – 1993 – zvýšení míry nezaměstnanosti z 3,5% koncem roku 1993 na 5–6% ke konci příštího roku*

III.4.2.2.2. Koreference deverbativ

Podobně jako v případě s abstrakty, reference a tudíž i koreference dějových jmen se jeví poměrně často velice problematickou. K referenci deverbativ jsou podstatné práce Krejdlin – Rachilinová (1981) a I. Mendozové (2004). I. Mendozová nahlíží na referenci deverbativ stejně jako na referenci jiných substantiv s několika omezeními na referenční schopnosti. Na základě referenční klasifikace předmětných jmen Padučevové (Padučevová 1979) autoři článku (Krejdlin – Rachilinová 1981) vyčleňují následující referenční typy deverbativ:

- specifická reference:

(76) rus. *Актриса с большим трудом привыкла к перемещению фотокамеры.*
č. *Herečka si jen těžko zvykla na posun kamery.*

- existenciální reference:

(77) rus. *Соссюр первый заметил связь этих двух явлений.*
č. *Saussure jako první všiml souvislosti těchto dvou jevů.*

- universální reference:

(78) rus. *Он выступил против преследования негров.*
č. *Veřejně vystoupil proti pronásledování černochů.*

- generická reference:

(79) rus. *Он выступил против преследования негров как типичной формы проявления расизма.*
č. *Promluvil proti pronásledování černochů jako typické formy rasismu.*

- hypotetická reference:

(80) rus. *Стрелочник предотвратил крушение поезда.*
č. *Výhybkář zabránil havárii vlaku.*

Tyto typy jsou založeny na významu dvou příznaků – existence do momentu *t* a relevantnosti protikladu singuláru a plurálu.

Zohlednění uvedených názorů nás vede k uznání možnosti specifické i nespecifické reference u deverbativ. K důslednému provedení toho rozlišení nás vedou i skutečná data z korpusu PDT. S deverbativy tedy zacházíme následujícím způsobem:

1. Jsou-li oba členy vztahu deverbativa s předmětným významem, anotujeme jejich vztah jako koreferenci u předmětných NP (viz III.4.2.1.1. – III.4.2.1.3.). Srov. příklad (81)a–b – návod na vyplnění daňového přiznání DPH.

(81) *a. Příslušnou rubriku najdete na 2. straně tiskopisu přiznání označenou jako položka 2 – Odpočet při změně režimu.*
b. Doklady k odpočtu se k přiznání {coref_text, typ=NR na "přiznání" v *a.*} *nepřikládají.*

2. Je-li druhý člen páru deverbativum s abstraktním (dějovým) významem, rozlišujeme mezi specifickou a generickou referencí a anotujeme vztah mezi deverbativem a antecedentem následujícím způsobem:

a) Pokud obě deverbativa mají specifickou referenci a jejich aktanty jsou koreferenční, nebo druhý člen páru anaforicky odkazuje na propoziční antecedent, označující konkrétní situaci, koreferenci anotujeme a označujeme ji jako typ 0. Do této skupiny patří většina odkazů na situaci (slovesnou frázi, propozici, větu). Srov. např. (82)a–b:

(82) *a. Malé a střední podniky zvyšují svůj podíl na vyrobeném produktu i na zaměstnanosti a jejich počet neustále roste.*
b. Tuto skutečnost {coref_text, typ=0 na "a" v *a.*} *jednoznačně konstatuje ministr hospodářství Karel Dyba v analýze, kterou předložil vládě.*

Srov. také příklad (83), s koreferenčním vztahem typ 0 mezi deverbativy se specifickou referencí:

(83) *a. Nová striktní omezení vlády SR proti českým exportérům*
[... 12 vět ...]
b. Již několik dnů je všeobecně známo, že ochranářská opatření {coref_text, typ=0 na "omezení" v *a.*} *slovenské vlády proti českým exportérům se dotýkají zejména oblasti obchodu s potravinami a zemědělskými produkty.*

b) Pokud obě deverbativa mají generickou referenci, resp. pokud jejich aktanty mají generickou platnost, anotujeme jejich koreferenci a označujeme ji jako typ=NR. Srov. generickou koreferenci NP *satelitního vysílání* a *toto bezdrátové spojení* v (84)a–b:

(84) *a. Zatím v ČR nikdo neodpověděl Českým radiokomunikacím na nabídku využít satelitního vysílání.*

[... 2 věty ...]

b. Některé banky už projevily zájem o toto bezdrátové spojení {coref_text, typ=NR na "vysílání" v *a.*}, *které umožňuje výměnu digitálních dat mezi svými pobočkami i s ostatním světem.*

Srov. také generickou koreferenci pro *vypořádání* v (85)a–b:

(85) *a. Rychlé, avšak i bezpečné vypořádání.*

b. Rychlost vypořádání burzovních obchodů v čase T+3 odpovídá podle Jiřího Béra, ředitele Burzovního registru cenných papírů při Burze cenných papírů Praha potřebám.

c) Pokud obě deverbativa mají specifickou referenci ale jejich aktanty nejsou koreferenční, koreferenční vztah mezi nimi neanotujeme.

d) Pokud jedno deverbativum má specifickou a druhé generickou referenci, koreferenční vztah mezi nimi neanotujeme. Srov. žádný vztah mezi *provoz* v (86)a a (86)b:

(86) *a. Linka 855 44 33 bude v provozu nepřetržitě 24 hodin.*

[... 8 vět ...]

b. V těchto obdobích bude provoz {žádný koreferenční vztah} *na Lince bezpečí zněkolikanásoben , objasňuje ředitelka linky.*

III.4.2.2.3. Koreference pojmenovaných entit

Samostatnou pozornost zaslouží koreferenční vztahy, kde aspoň jeden z členů páru je pojmenovanou entitou.[102] Anotaci koreference pojmenovaných entit je věnována v současné počítačové lingvistice velká pozornost. Informace o vztazích pojmenovaných entit v souvislém textu jsou velmi důležité pro řešení automatické extrakce informace z textu a mnoha jiných aplikačních úkolů. Kromě toho, v současné době již existují fungující programy automatického rozpoznávání pojmenovaných entit (Named Entities Workshop v Singapuru: Haizhou – Kumaran 2009; Sekine, 2004; Collins – Singer 1999; Talukdar a kol., 2006; Santos a kol. 2006, Sassano – Utsuro 2000 a mnoho dalších). Je to velmi důležitý krok pro zpracování koreference, protože anotace koreference na omezené a přesně definované množině výrazů bude mít přesnější, a tedy použitelnější výsledky.

Na našem pracovišti ÚFAL v současné době skupina počítačových lingvistů pracuje na automatickém rozpoznávání a klasifikaci pojmenovaných entit (Ševčíková a kol. 2007a a 2007b, Kravalová – Žabokrtský 2009). Teprve vznikající informace však zatím není možné použít pro naši koreferenční analýzu. V budoucnu se počítá s jejich propojením.

Anaforické vztahy mezi pojmenovanými entitami, stejně jako v případě jiných typů výrazů, se dělí na textovou koreferenci a asociační anaforu. Rozlišujeme tedy vztahy *Bělorusko – president* (asociační anafora) a *Lukašenko – president* (identická koreference).

Textová koreference a asociační anafora se anotuje rovněž u odvozených od pojmenovacích entit adjektiv (viz III.2.1.2.), která se chápou jako koreferenční se substantivy, od kterých jsou odvozena. Toto pravidlo je původně formulováno pro všechny pojmenované entity, ale de facto se jedná téměř výhradně o pojmenovaných entitách odvozených od geografických názvů (*Německo – německý* apod.). Textovou koreferencí jsou tedy propojeny páry typu:

- substantivum – substantivum (*Německo – Německo*),
- substantivum – adjektivum (*Německo – německý*),
- adjektivum – substantivum (*německý – Německo*),
- adjektivum – adjektivum (*německý – německý*).

Pro zpracování koreferenčního vztahu mezi pojmenovanými entitami označujícími geografické názvy, jejich zkratkami a odvozenými adjektivy používáme automatickou předanotaci (viz IV.1.2.1.).

Z formálního hlediska koreference pojmenovaných entit vypadá takto:

[102] Pojmenovaná entita = vlastní jméno, proprium.

1. Velká část případů identické textové substantivní koreference je opakování stejného výrazu, přičemž často anafor a antecedent jsou identické – opakuje se celý podstrom, téměř nikdy (pokud jde o spisovnou češtinu a texty PDT) se anafor nepoužívá s ukazovacím zájmenem. Srov.:

(87) *a. Pouze z bývalé Šternberské konírny v přízemí křídla přiléhajícího k Thunovské uličce se stane konferenční (tiskový) sál.*

[... 17 vět ...]

b. Když architekti zvažovali optimální propojení staré budovy sněmovny s novými domy , vsadili na tunel pod Thunovskou uličkou.

2. Další podstatná skupina koreference pojmenovaných entit je koreference pojmenované entity a obecného jména. Můžeme vyčlenit následující frekventované skupiny:

a) anafor je název funkce osobního antecedentu (typ *Lukašenko – president, Karel Dyba – ministr* apod.), ale zůstává přitom identická koreference;

b) antecedentem je pojmenována entita, anaforem je obecné jméno při té pojmenované entitě, která na něm visí jako přímý potomek s funktorem ID nebo RSTR, tj. anafor je např. *firma Struktura, země Španelsko, v Sekaninově ulici, projekt Světlo v temnotách* apod.). V tom případě vzniká otázka, který uzel má koreferovat – pojmenovaná entita nebo obecné jméno. Z hlediska struktury stromu a pravidel anotace jiných úseků (viz princip maximální velikosti koreferujících členů v III.1.3.) vyplývá rozhodnutí koreferovat vždy na formálně řídící uzel (tj. na *firma, země, ulice, projekt* apod.) Srov. (88)a–f:

(88) *a. V Košťanech totiž zakoupila dům firma Struktura, která* {coref_gram na „firma"} *se u nás rozmísťováním německých chlapců zabývá.*

b. Posledním místem, kam byli chlapci firmou {coref_text, typ=0 na „který" v *a.*} *Struktura umístěni , byl bývalý dům dětí a mládeže v Duchcově .*

c. Tajemná Struktura {coref_text, typ=0 na „firma" v *b.*}

d. Ten, kdo ve skutečnosti německé chlapce v severočeském pohraničí umísťoval [], byla firma {coref_text, typ=0 na „Struktura" v *c.*} *Struktura s.r.o. [], která* {coref_gram na „firma"} *se zabývá sociálním managementem.*

e. Její {coref_text, typ=0 na „který" v *d.*} *zástupce ing. Šedivý však veškerou*

odpovědnost za krizovou situaci odmítá.

f. Mezitím starosta obce Košťany Jindřich Abrhám požádal o pomoc člena branně – bezpečnostního výboru parlamentu poslance Čapka (Levý blok), který se obrátil na ministra Rumla s žádostí o prošetření činnosti firmy {coref_text, typ=0 na `#PersPron` *(její)* v *e.*} *Struktura.*

Problém vzniká v případě, že řídící obecná jména při pojmenované entitě s funktorem ID nejsou koreferenční. V tom případě nemůžeme mezi nimi navázat koreferenční vztah a pojmenované entity zůstávají nepropojené. Srov. *systémem A – Consult plus* v (89)a–b:

(89) *a. Oceňování hmotného investičního majetku systémem A – Consult plus – týdenní profesní školení, které je určeno pro přípravu soudních znalců pro oceňování hmotného investičního majetku, zvláště pro účely úvěrového řízení v bankách a ve všech případech, kdy není zapotřebí cenový předpis.*

b. Absolvent získá licenci k používání softwaru znaleckého ústavu A – Consult plus {žádný koreferenční vztah} *a následný servis na jeden rok.*

Srov. také *středoevropský – Střední Evropa* v (90), kde řídící uzel *Evropa* má funktor `ID` a tedy nemůže být propojen s adjektivem *středoevropský*; celá jmenná fráze *názvem Střední Evropa* však už se *středoevropský* koreferenční není:

(90) *Španělská sekce prestižní Asociace evropských novinářů pozvala již po šesté různorodé spektrum středoevropských politiků, publicistů a filosofů, aby početným posluchačům letního universitního semináře s názvem Střední Evropa.*`ID` {žádný koreferenční vztah} *mezi Bruselem a Moskvou přednesli své úvahy o tom, co se uprostřed kontinentu děje.*

III.4.2.2.3.1. Anotace víceslovných pojmenovaných entit

V případě, že víceslovná pojmenovaná entita, referuje k jednomu objektu reality a její součástí nejsou jiné pojmenované entity referující k jiným objektům reality (viz III.4.2.2.3.2.), za koreferující člen se považuje formálně řídící uzel. U závislých uzlů se koreference nezaznamenává. Srov. příklady v tabulce č. 18 (koreference se anotuje pouze u podtržených uzlů):

České Budějovice – České Budějovice

Česko – Česká Republika – ČR

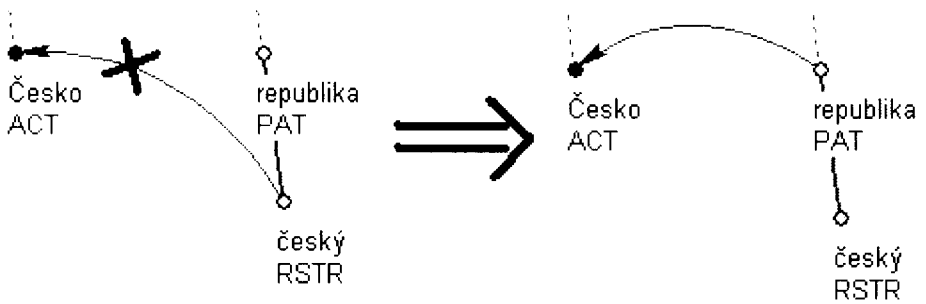
Václav Havel – prezident Havel

ministr hospodářství Karel Dyba – ministr Karel Dyba – Karel Dyba

prof. PhDr. Zdeněk Matějček – Zdeněk Matějček

Tabulka č. 18: Anotace víceslovných pojmenovaných entit

Se zřetelem k uvedenému pravidlu má být kontrolována a opravena automatická předanotace koreference geografických názvů (viz IV.1.2.1.) Srov. typický příklad chyby na obrázku č. 15:



Obrázek č. 15: Opravování koreference u adjektiv odvozených od pojmenovaných entit

V určitých kontextech jsou možné další případy, kde původní automatická anotace koreference pojmenovaných entit má být opravena např. *Malta* a *maltský tým* v (91)a a (91)b, kde „Malta" referuje k týmu, nikoliv k Maltě jako ke státu). V (91)b se adjektivum *maltský* nepropojí s ničím, protože k týmu referuje řídící uzel a Malta-země tam není. Ale propojilo by se s Maltou třeba ve větě „čeští fotbalisté odletěli na Maltu" paralelně vedle řetězce referujícího k týmu.

(91) *a. S Kadlecem, či bez něho – to je klíčová otázka trenéra české fotbalové reprezentace Dušana Uhrina, kterou musí vyřešit před dnešním úvodním utkáním kvalifikace ME v 16.30 v Ostravě s Maltou.*

b. V maltském {žádný koreferenční vztah} *týmu* {coref_text, typ=0 na „Malta"} *jsou dva pamětníci kvalifikace MS 1986, kdy čs. reprezentace na jejich hřišti ztratila bod po bezbrankové remíze – obránci Buttigieg a zvláště donedávna jediný krajánek Busuttil.*

III.4.2.2.3.2. Anotace částí pojmenovaných entit

Jmenná skupina víceslovné pojmenované entity může obsahovat další výrazy, které se potom opakují v následujícím textu samostatně. Srov. například NP *výzkum* a *rodina* v (92)b–f, které jsou součástí pojmenované entity *Oddělení pro výzkum rodiny* v (92)a:

(92) *a. ... prof. PhDr. Zdeněk Matějček a Doc. MUDr. Zdeněk Dytrych před pětadvaceti lety založili Oddělení pro výzkum rodiny, které dodnes vedou.*

b. Z. Matějček se věnuje dětem a Z. Dytrych dospělé části rodiny {žádný koreferenční vztah na „Oddělení pro výzkum rodiny" v *a.*}.

[... 16 vět...]

c. Nebo například existuje lehká mozková disfunkce, kterou trpí podle našeho rozsáhlého výzkumu {žádný koreferenční vztah na „Oddělení pro výzkum rodiny" v *a.*} *pět procent dětí.*

[... 17 vět...]

d. Materiálům, které dnes máte k dispozici, předcházel dlouholetý výzkum {žádný koreferenční vztah na „Oddělení pro výzkum rodiny" v *a.*}.

e. Zdeněk Dytrych: Od roku 1969, kdy jsme založili v bývalém Výzkumném ústavu psychiatrickém Oddělení pro výzkum rodiny {coref_text, typ=0 na „Oddělení pro výzkum rodiny" v *d.*} *, se hlavně zabýváme touto problematikou.*

f. Tak například rozsáhlý výzkum {žádný koreferenční vztah na „Oddělení pro výzkum rodiny" v *d.*} *rozvodovosti.*

Zatímco mezi *výzkum* v (92)a a *výzkum* v (92)c,d,f není vztah textové koreference (výzkum v (92)a má generickou referenci, v (92)c,d,f je možné hovořit o specifické referenci), jediný vztah, o kterém v daném případě můžeme uvažovat, je asociační anafora typu „část – celek", v případě NP *rodina* v (92)a a (92)b; v obou větách je generická reference. Nejde tu však o koreferenci ani v nejširším smyslu, jak jsme ji definovali v úvodu této práce. Podle našeho názoru, obecná jména, která jsou součástí pojmenovaných entit, se chovají z referenčního hlediska podobně jako uzly s funktorem `ID`, tj. nejsou plnohodnotně referenční. Proto považujeme za vhodné neanotovat koreferenční vztahy u obecných názvů, které jsou součástí pojmenovaných entit.

Jiná situace nastává v případě, že součástí pojmenované entity je jiná pojmenovaná entita, referující k dalšímu objektu reality. V textech PDT to je především případ, kdy vložená pojmenovaná entita je geografický název. Srov. v (93)a–c:

(93) *a. Našemu listu se podařilo získat od představitelů ÚNMS SR*[103] *informaci o technickém zajištění propouštění potravinářských výrobků do SR.* {coref_text, typ=0 na „SR" v *a.*}

b. Z rozhodujících opatření, která by měla plně vstoupit v platnost po 1. dubnu vyjímáme:

c. Do 31. března platí v plném rozsahu postup podle dohody ÚNMZ ČR a ÚNMS SR {coref_text, typ=0 od „SR" na „SR" v *b.*}, *na jejímž základě český výrobce (slovenský* {coref_text, typ=0 na „SR" v *c.*} *dovozce) získá na základě schválení české zkušebny a rozhodnutí Ministerstva zdravotnictví SR* {coref_text, typ=0 na „slovenský"}*na ÚNMS SR* {coref_text, typ=0 na „SR" v *c.*} *potvrzení o platnosti rozhodnutí i na území SR* {coref_text, typ=0 na „SR" v *c.*}.

V případě reference pojmenovaných entit, jde o referenci jedinečnou a tedy vždy specifickou a určitou, tj. všechna pojmenování označující daný objekt jsou koreferenční. Proto pokládáme za smysluplné anotovat koreferenci u vložených pojmenovaných entit. Také z technických důvodů je jednodušší je propojit, než je nechat nepropojené (tj. de facto bychom museli vždy ručně smazat automaticky předanotovanou koreferenci u všech vložených pojmenovaných entit, což je pracné a zbytečné). Kromě toho jednotná anotace pojmenovaných entit je jedinou možnou variantou případné budoucí automatické anotace zaměřené na řešení úloh počítačové lingvistiky (strojové učení, automatická excerpce informace, strojový překlad aj.). V příkladě (93) tedy zaznamenáváme textovou koreferenci mezi ÚNMS SR v (93)a a (93)c a zvlášť propojíme koreferencí všechny SR a adjektivum *slovenský* v (93)a–c.

Z uvedených argumentů plyne následující pravidlo:

Pravidlo anotace koreference u částí pojmenovaných entit:

Části pojmenovaných entit anotujeme pouze v případě, když jsou samy pojmenovanou entitou a referují k jinému objektu reality než jejich řídící uzel.

[103] Úřad pro normalizaci, metrologii a zkušebnictví SR.

Srov. také příklady v tabulce č. 19:

anotovat	*Ústavní soud ČR – ČR;*
	ÚNMS SR – SR.
neanotovat	*Oddělení pro výzkum rodiny – výzkum rodiny;*
	Ministerstvo zemědělství – zemědělství.

Tabulka č. 19: Anotace částí pojmenovaných entit

III.4.2.3. Problematické případy označování textové koreference

Typologie textově koreferenčních vztahů, která byla zavedena v III.4.2.1., je z hlediska struktury a koherence textu velice vágní, spíše schematická než vyčerpávající. Pro naši anotaci jsme určili hranici mezi specifickou a nespecifickou referencí, naznačili pravidla vymezení generických a specifických NP a stručně představili abstraktní a dějová substantiva, u kterých jsme se také rozhodli pro anotaci identické textové koreference. Je však zřejmé, že s těmito jevy nevystačíme při komplexním popisu koherence textu, ani při popisu identické koreference v textu, která k této koherenci přispívá. V anotaci označujeme jenom malou část vztahů, které vytváří jeho kohezi. Kromě toho, naše orientace na koreferenci vůči anafoře (viz princip preference koreference v III.1.6.) nás odvádí ještě dále od zaznamenávání skutečné textové koherence směrem k označování jednoho formálně a konvenčně definovaného jevu, který v podstatě ani nemusí mít se skutečnou organizací textu mnoho společného. Skutečná koherence textu však těžko může být jen zaznamenána, a už vůbec ne pomocí formálních metod a pravidel, která se dají reálně aplikovat na byť ruční ale jednotnou anotaci na velkém korpusu. Existují některé projekty, které se o to pokoušejí,[104] mají však buď experimentální povahu (na minimálním počtu textů, na malých vzorcích) a/nebo také zdaleka nejsou vyčerpávající. Schematicky by se koherence textu mohla představit jako velmi složitá síť vztahů různé kontextové váhy a délky, přičemž elementy a skupiny elementů mohly být propojeny mezi sebou více vztahy najednou a počet vztahů vytváří otevřenou množinu, která také jen stěží podléhá klasifikaci. Pokud jde konkrétně o koreferenci, předpokládáme, že existence koreferenčních řetězců v textu je pouze arbitrární, protože primární je vztah anaforický a v něm žádné řetězce nejsou – v anaforických vztazích se lehce prolínají elementy různých úrovní, kategorií a typu reference (*děti* (generická reference)– *moje děti* (specifická reference) – *děti v dětských domovech* (gen) – *dětské domovy* (vložená NP generická nebo specifická reference) – viz IV.4. – apod.). Koreference je v podstatě pouze formální analýza vztahů elementů v textu k

[104] Srov. např. u nás Novák a kol. 2009, Novák 2008, Novák a Hall 2008 aj.

objektům skutečnosti a může být představena i jiným způsobem, např. nasměrováním všech koreferenčních výrazů k prvnímu antecedentu nebo k rekonstruovanému fiktivnímu objektu skutečnosti.

Nicméně ve stávající anotaci textové koreference se držíme uvedených v III.4.2.1. typologických principů. Problematické případy, které z právě uvedených důvodů nelze do naší klasifikace jednoznačně začlenit, rozebíráme v následující kapitole.

III.4.2.3.1. Hraniční případy mezi [coref_text, typ=0] a [coref_text, typ=NR]

V některých případech u výrazů s předmětným významem je těžko odlišit specifickou referenci od generické. Většinou jsou možné obě interpretace. V současné době nejsme schopni tyto situace přesně vymezit a popsat, proto uvedeme jen několik příkladů ((94)–(97)), které znázorňují konkrétní rozhodnutí v jednotlivých situacích.

(94) *a. Po vojně začal v Masokombinátu v Ostravě – Martinově.*

[... 7 vět ...]

b. Ve státním podniku mne ubíjel stereotyp a nepružnost.

[... 14 vět ...]

c. „A samotný start po odchodu z Martinova"?

[... 19 vět ...]

d. Klidně jsem mohl seskočit a dál dělat ve státním podniku, nic by se nestalo.

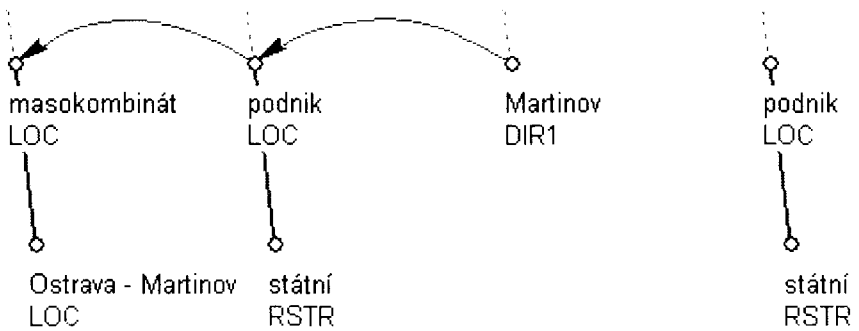
V uvedených větách (94)a–d nejde o anaforické odkazování, avšak pro koherenci textu je spojení podtržených jmenných frází poměrně důležité. Zamyslíme se nad referencí těchto NP v (94)a–d:

a) NP *Masokombinát v Ostravě – Martinově* v (94)a má jednoznačně specifickou referenci.
b) NP *státní podnik* v (94)b má dvojí referenční interpretaci – generickou (*ve kterémkoliv státním podniku; v tom, co je státní podnik; v takovém podniku, který je státní; ve státním podniku jako formě podniků* a jiné perifráze) a specifickou (v tom státním podniku, ve kterém jsem pracoval, čili v Masokombinátu v Ostravě – Martinově). Pravděpodobnější je asi generická interpretace, ale specifická také není vyloučena.
c) NP *Martinov* v (94)c má také dvě referenční interpretace, ale obě jsou specifické – v prvním případě *Martinov* referuje na město, ve kterém pracoval adresát otázky (94)c, ve

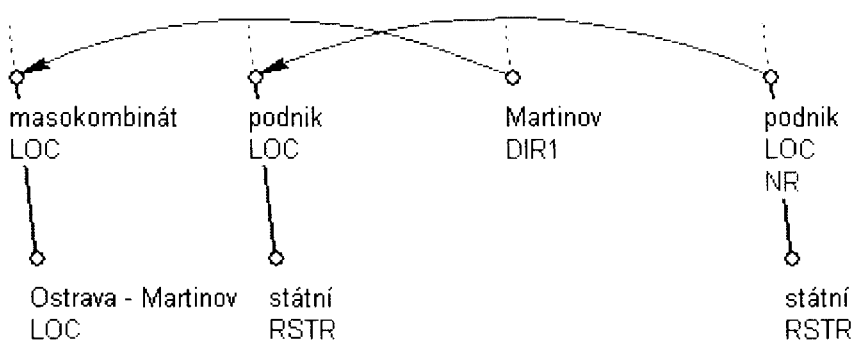
druhém případě *Martinov* referuje metonymicky na podnik, tj. na Masokombinát v Ostravě – Martinově.

d) NP *státní podnik* v (94)d má s největší pravděpodobností generickou interpretaci.

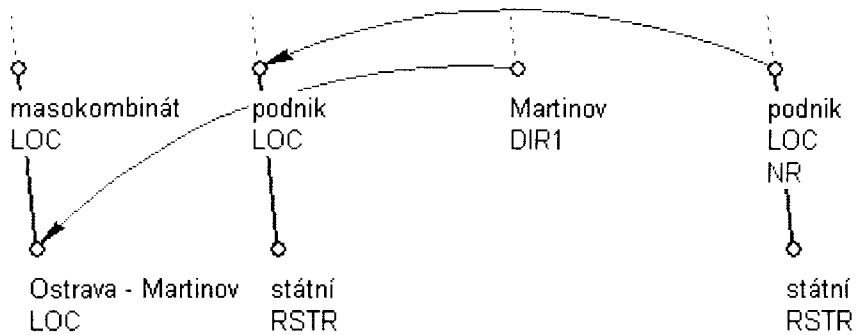
Z různých referenčních interpretací v (94)b–c plyne více možností navázání koreferenčního vztahů mezi těmito výrazy – srov. několik možností označení koreference, schématicky zobrazených na obrázcích č. 16–19.



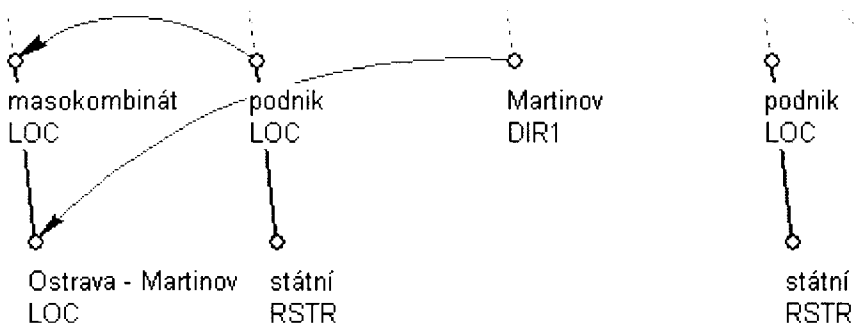
Obrázek č. 16: Nejednoznačnost koreferenčních vztahů. „Podnik" má specifickou referenci, „Martinov" je chápáno metonymicky



Obrázek č. 17: Nejednoznačnost koreferenčních vztahů. „Podnik" má generickou referenci, „Martinov" je chápáno metonymicky



Obrázek č. 18: Nejednoznačnost koreferenčních vztahů. „Podnik" má generickou referenci, „Martinov" odkazuje na město



Obrázek č. 19: Nejednoznačnost koreferenčních vztahů. „Podnik" má specifickou referenci, „Martinov" odkazuje na město

Každá z uvedených interpretací je oprávněná. Při rozhodnutí se snažíme dodržovat principy anotace (viz III.1.), ale v komplikovaných příkladech to není vždy jednoznačně. V daném příkladě jsme se rozhodli pro variantu z prvního obrázku č. 16 na základě principu dodržování maximálního koreferenčního řetězce. NP *státní podnik* v (94)d pak bude propojena s posledním uzlem daného koreferenčního řetězce asociační anaforou typu „část – celek" (viz III.5.1.1.).

Srov. další příklad (95)a–b:

(95) *a. U detergentu Toto jsme například řešili problém s udržením stálé kvality, protože jednotlivé partie byly nevyvážené.*

b. Investovali jsme dva miliony korun do nákupu pásových vah, zpřesnili

dávkování a jakost pracího prášku stabilizovali.

V (95)a–b rozhodujeme mezi generickou a specifickou referencí NP *detergentu Toto* a *pracího prášku*. Jmenné fráze jsou v tomto příkladě jednoznačně koreferenční – odkazují ke stejné značce pracího prášku. Existuje však možnost specifické a generické interpretace jejich reference:

a) obě NP v (95)a a (95)b mají generickou referenci, tj. referují na typ pracího prášku – nikoliv na konkrétní značku, ale na celý druh, na prototypického představitele dané třídy. V tom případě vztah mezi nimi je označen jako coref_text, typ=NR.
b) obě NP v (95)a a (95)b mají specifickou referenci, tj. referují na vybranou a jedinečnou značku pracího prášku. V tom případě vztah mezi nimi je označen jako coref_text, typ=0.

Na základě pravidla o preferenci defaultní textové koreference, typu 0 (viz III.4.2.1.1.) a také proto, že jde o pojmenovanou entitu, o kterých víme, že v kontextu mají výraznou tendenci dostávat specifickou referenci, volíme variantu (b) se specifickou referencí.

Daný příklad nás zavádí do ještě jedné problematické oblasti, a to nakolik specifická má být specifická reference. Úvahy na dané téma najdeme již u Fregeho (Frege 1892) a dále skoro ve všech pracích věnovaných teorii reference (viz odkazy v III.2.). Jde o rozdíl mezi reálným objektem mimojazykové skutečnosti (denotát/referent) a jeho myšlenkovým obrazem, který se vytváří v naší mysli (referent/denotát). Specifická reference je určována na základě myšlenkového obrazu nikoliv skutečné existence objektu, proto např. pohádkové postavy a neexistující zvířata mohou mít specifickou referenci. Tuto myšlenku můžeme rozšířit i na jiné oblasti než předmětná jména. Tím chceme říci, že pokud máme o objektu v daném kontextu jedinečný a přesně definovaný myšlenkový obraz, můžeme hovořit o jeho specifické referenci, byť nemá v reálném světě materiální referent (denotát), na který můžeme sáhnout, ale označuje například příznak, situaci apod.

Srov. také podobný příklad (96), kde NP *tento poplatek* v (96)b nemá materiální denotát. Stačí to pro to, aby daná jmenná fráze byla interpretována jako generická? Momentálně tento typ označujeme jako textovou koreferenci NP se specifickou referencí (coref_text, typ=0):

(96) *a. Milionový poplatek za vydání osvědčení, které umožňuje vést lékárnu, zakázalo vybírat Ministerstvo pro hospodářskou soutěž.*

b. Tento poplatek {coref_text, typ=0 na „poplatek" v *a.*} *odhlasovali její členové na svém druhém sjezdu v říjnu 1992.*

Srov. další příklad (97))a–b:

(97) *a. Začal jsem provozováním hospody, která* {coref_gram, na „hospoda"} *byla mnohokrát vykradena.*

[... 2 věty ...]

b. Hospoda {coref_text, typ=0 na „který" v *a.*} *byla jen startem, polem k podnikání s masem a masnými výrobky.*

V (97))b jsou možné dvě různé referenční interpretace NP *hospoda:*

a) NP *hospoda* v (97)b má specifickou referenci (ta konkrétní hospoda, kterou podnikatel provozoval). V tom případě vztah mezi *hospoda* v (97)a a *hospoda* v (97)b je coref_text, typ=0.

b) NP *hospoda* v (97)b má generickou referenci (hospoda jako taková – podnikatel chtěl pořídit něco jako hospodu, aby poznal svět podnikání). V tom případě vztah mezi *hospoda* v (97)a a *hospoda* v (97)b je coref_text, typ=NR.

Takových nejednoznačných referencí je nečekaně mnoho. Pokud není úplně zřetelná nereferenční interpretace, budeme je na základě pravidla o preferenci specifické reference (viz III.4.2.1.) anotovat jako vztah mezi NP se specifickou referencí.

III.4.2.3.2. Hraniční případy mezi [coref_text, typ=NR] a vztahy, které lze chápat jako nekoreferenční

Koreference jmenných frází s nespecifickou referencí má velmi nepřesnou hranici mezi tím, kde koreference ještě má být zaznamenána a tím, kde o koreferenci už nejde ani v tom nejširším smyslu, jak jsme ji definovali v úvodu. Vágnost přechodu je podmíněna především tím, že koreference generických NP není skutečnou koreferencí dvou objektů (viz diskuzi na

toto téma v III.4.2.1.). Při rozhodování mezi zaznamenáním a nezaznamenáním generické koreference se řídíme následujícím negativně formulovaným pravidlem:

Pravidlo o (ne)zaznamenání generické koreference:

Generickou koreferenci neanotujeme v následujících případech:

1. Nereferenční NP mají různý dosah/extenzi (tj. vztahují se na různé množiny objektů, např. *žena – žena v 19. století*);

2. Splňují se obě následující podmínky:

a) nejsme si jisti koreferencí daných NP;

b) vztah mezi danými NP nemá vliv na kohezi textu.

V ostatních případech generickou koreferenci v podobných problematických případech anotujeme.

Případů z PDT, kde nemůžeme na základě definovaných kritérií rozhodnout o anotaci koreference mezi generickými NP, je velký počet. Podobně jako v předchozí kapitole uvedeme zde několik nejtypičtějších příkladů s vysvětlením každého konkrétního rozhodnutí.

V (98)b koreferenci u NP *podnikatelé* neanotujeme. S největší pravděpodobností množiny podnikatelů, na které referují jmenné fráze v (98)a a (98)b nejsou identické. Vztah mezi *podnikatelé* v (98)a a (98)b nemá přímý vliv na kohezi textu.

(98) *a. Mnozí čeští podnikatelé si ke své škodě stále ještě neuvědomují , že peníze věnované na získání důležitých údajů se jim mnohonásobně vrátí.*

[... několik vět ...]

b. Naším cílem je , aby podnikatelé {žádný koreferenční vztah}věděli o sobě navzájem.

Srov. následující příklad (99)a–c:

(99) *a. Kdy děti nejvíce volají*

[... několik vět ...]

b. Pražské děti budou mít hovor {žádný koreferenční vztah} *zdarma.*

c Podle zkušeností ze zahraničí se dá předpokládat, že největší frekvence

telefonátů {coref_text, typ=NR na „volat" v *a.*} *nastane vždy mezi 16. až 18. hodinou.*

Zde v (99)b je deverbativum *hovor* s generickou (příp. distributivní) referencí na množinu případných telefonátu z Prahy. Proto (99)b nemůže odkazovat na situaci v (99)a, kde se mluví o telefonátech všech dětí, nejenom z Prahy. Naopak, NP *telefonáty* v (99)c zřejmě genericky odkazuje na všechna možná volání dětí a je v podstatě substantivizací situace s voláním v (99)a. Přestože generické odkazování na situaci je poněkud marginální, v daném případě můžeme *volat* v (99)a a *telefonát* v (99)c propojit generickou textovou koreferencí.

Podobný předchozímu (99) je následující příklad (100)a–c, avšak v daném případě jsme vztah mezi deverbativem *rozvoz* v (100)c a *doprava* v (100)b nezaznamenali. Je to vztah těžko zachytitelný a nezdá se, že by mohl mít pro strukturu textu důležitý význam. Případná souvislost je již zaznamenána v anotaci aktuálního členění (*rozvoz* bude v topiku):

(100) *a. Po vojně začal v Masokombinátu v Ostravě – Martinově.*

b. V dopravě, ale zajímal se o všechno z provozu.

[... 9 vět ...]

c. Třeba při rozvozu {žádný koreferenční vztah} *jsem denně přenesl pěkných pár tun na zádech.*

Další příklad (101)a–c je problematičtější: intuitivně *problémek* a *problém* v (101)c nejsou koreferenční s výrazem *problém* v (101)a. V obou případech však jde o generickou referenci na problémy dětí, kteří volají na linku důvěry. V daném příkladě jsme koreferenční vztah nezaznamenali – zdá se, že *problém* v (101)c má spíše predikativní nereferenční interpretaci (podobně jako se slovesem *být* v predikativních konstrukcích), zatímco *problémek* referuje k nějakému menšímu problému, tedy ne ke všem problémům, se kterými volají děti v (101)a.

(101) *a. Chceme, aby ze sebe problémy dostaly ven.*

b. Jakmile se začnou svěřovat, už se s tím dá něco dělat.

c. Jde o to, aby se z problémku {žádný koreferenční vztah} *nestal problém* {žádný koreferenční vztah}.

Srov. také vztah mezi *paláce* a *klasické renesanční a barokní paláce* (102)a–c:

(102) *a. Svědkem oněch časů zůstal mj. i pseudorenesanční spojovací můstek mezi sněmovnou a Šternberským palácem z roku 1910.*

b. Paláce {žádný koreferenční vztah} *neznamenají přepych.*

c. Ač se to na první pohled nezdá, obývání klasických renesančních a barokních paláců {coref_text, typ=NR na „palác" v *b.*} *s velikými, řetězovitě propojenými místnostmi není žádné terno.*

V páru *palác – palác* (102)a a (102)b textovou koreferenci neoznačujeme, protože v (102)a výraz *palác* má specifickou referenci, zatímco v (102)b – generickou, tedy (102)a a (102)b nejsou koreferenční. Vztah mezi *paláce* a *klasické renesanční a barokní paláce* v (102)b a (102)c je zajímavější. Obě jmenné fráze mají generickou referenci, ale na první pohled referenční oblast druhé NP je vložena do referenční oblasti NP *paláce.* Avšak v daném kontextu i výraz *paláce* v (102)b referuje k množině klasických a renesančních paláců – je to podtitul a žádné jiné paláce se v jeho dosahu neobjevují. Proto považujeme za smysluplné označit mezi podtrženými NP v (102)b a (102)c textovou koreferenci s typem NR.

V následujícím příkladu (103)a–b textovou koreferenci typu NR neoznačíme, protože reference výrazů *kvalita* v (103)a a (103)b mají různý dosah, *kvalita* v (103)b je „specifičtější" než kvalita v (103)a (jde o kvalitu pouze detergentu), tj. odkazuje na omezenější množinu, třídu denotátu.

(103) *a. Stali jsme se také dodavatelem Unileveru a dokázali splnit jeho zvýšené požadavky na kvalitu.*

b. U detergentu Toto jsme například řešili problém s udržením stálé kvality {tady – kvality pouze detergentu proto žádný koreferenční vztah}, *protože jednotlivé partie byly nevyvážené.*

III.4.2.3.3. Spojení s výrazy s významem „kontejneru"

Spojení s výrazy s významem „kontejneru" v širokém smyslu slova (např. *sklenice, láhev, sud,* ale také *dostatek, spousta, počet, dávka, skupina, polovina, balení, část, stádo, většina, řada* a číslovek *tisíc, milion, miliarda* apod.)[105] bývá při anotaci rozšířené textové koreference

[105] K výrazům s významem „kontejnerů" a jejich anotaci na tektogramatické rovině v PDT viz (Mikulová a kol. 2005).

rovněž problematickou záležitostí. V anotaci původní zájmenné koreference za antecedent se počítala ta jednotka, se kterou se shodovalo anaforické zájmeno. Srov. koreference na kontejner v (104) a na závislý člen v (105)a–b:

(104) *Absolutní většina lidí závislých na heroinu je příliš mladá na to, aby si #PersPron* {coref_text na „většina", nikoliv na „lidé"} *pamatovala rozklad a zesláblost generace sedmdesátých let, takže odvrácenou stránku „fantastického" života si mnohdy vůbec neuvědomí.*

(105) *a. Předcházet bude řada postupných kroků.*
b. Jedním z nich {coref_text na „krok", nikoliv na „řada"} *bylo právě navrhované navýšení o 150 až 300 mil. Kč, kvůli kterému byla přerušena mimořádná valná hromada.*

V případech bez gramatické shody koreference se anotovala na kontejner. Srov. např. v (106)a–b – (108) odkazování pomocí ukazovacího zájmena *to*, které nevyjadřuje gramatické příznaky antecedentu:

(106) *a. Ale přitom hostitel otevíral láhve alkoholu.*
b. Byla to {coref_text na „láhev", nikoliv na „alkohol"} *vína a německé alkoholy tvrdé.*

(107) *Společnost zaměstnává přes dva tisíce zaměstnanců, to* {coref_text na „tisíc", nikoliv na „zaměstnanec"} *je po propuštění důchodců a brigádníků prakticky stejně jako před pěti lety.*

(108) *Není potom divu, že ze zahraničí bylo za posledních pět let v Rusku investováno jen 2,7 miliardy dolarů (v roce 1994 se očekává, že k tomu* {coref_text na „miliard", nikoliv na „dolar"} *přibude další jedna miliarda), zatímco v Maďarsku, které má patnáctkrát méně obyvatel, to za stejné období bylo šest miliard USD.*

Při anotaci rozšířené textové koreference se můžeme setkat s řetězci, kde se střídá pronominální a rozšířená koreference. Abychom dodrželi jednotnost anotace a princip maximální délky koreferenčních řetězců (III.1.3.), musíme se držet stejného principu, tj.

anotovat koreferenci na kontejner. Srov. v (109)c budeme chtít spojit textovou koreferencí *osoba* s *jejich* v (109)b, tedy i s *desítka* v (109)a, nikoliv pouze s *osoba* v (109)a:

(109) *a. Na Chalupův návrh StB totiž zorganizovala asi dvěma desítkám osob přechody přes hranice.*
b. Chalupa slíbil, že s jejich {coref_text na „desítka", nikoliv na „osoba" v *a.*} *pomocí vybuduje později v exilu organizaci, která bude – ovšemže nevědomky – sbírat informace pro StB (nakonec se Chalupa angažoval zejména v Československém zahraničním ústavu v exilu).*
c. Mezi osoby {coref_text na „jejich"}, *které byly přes hranici převedeny přímo pracovníkem StB, patřil právě i Mojmír Povolný.*

Srov. také koreferenci na kontejner v následujícím příkladě č. (110):

(110) *a. Tří a půl tisíce dělníků vyhlásili stávku.*
b. Stávkující {coref_text na „tisíc" v *a.*} *žádají zvýšení platů o šest procent.*
c. Do 8. března se počet {coref_text od „počet" na „stávkující" v *b.*} *stávkujících může zdvojnásobit.*

Avšak u substantivní koreference, pokud přitom ani nejde tolik o anaforický vztah, gramatická shoda nehraje většinou žádnou roli (shoda v rodě dvou různých substantiv je vždy nahodilá). Srov. různý gramatický rod koreferenčních substantiv se specifickou referencí *hmota* a *prášek* a stejný rod (maskulinum) kontejneru *gram* a anafora *prášek* v (111):

(111) *V koupelně jednoho baru na newyorské Lower East Street rozmíchá žiletkou asi gram bílé hmoty a prášek* {coref_text na „gram"} *nasypaný na zrcadlovou plochu pouzdra nabízí svému méně zkušenému příteli.*

Kromě toho uzly, které jsou závislé na kontejnerech, se mohou zúčastnit jiných koreferenčních vztahů. Srov. v (112)a–f kontejner *počet* se závislým uzlem *celek* vystupuje v textu dohromady v (112)a,c,d,e, ale *celek* se používá také zvlášť v (112)b,f, přičemž *celek* v (112)b,f je koreferenční s *celek* v (112)a,c,d,e. V takovém případě nabízíme anotovat oba koreferenční řetězce – u kontejnerů a u jejich závislých uzlů.

(112) *a. Ve srovnání s vládní bitvou o počet celků z konce června nesla obě tato jednání*

znaky naprosté selanky.

b. Václav Klaus sice opakovaně ubezpečuje, že vláda kompetence <u>*celků*</u> {coref_text na „celek" v *a.*} *považuje za důležitější než jejich množství a vymezení, jeho slova však zanikají ve hřmotu návrhů na nejrůznější a často též nejnepochopitelnější počet celků a vzájemného obviňování se ze lži a porušování předchozích dohod.*

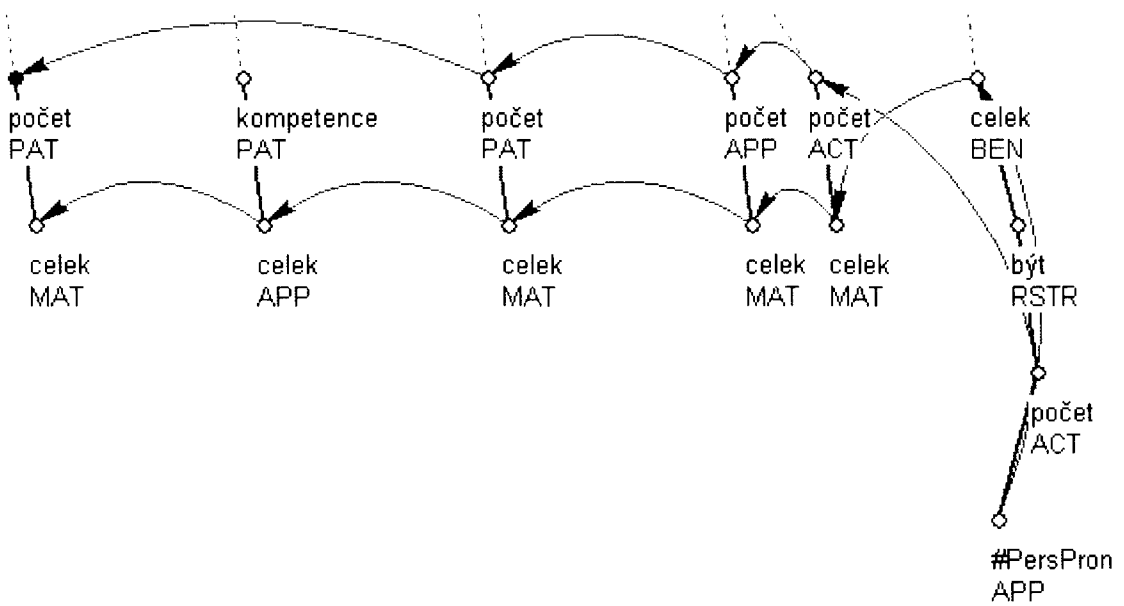
c. Pravděpodobnější se proto zdá být opak: právě <u>*počet*</u> {coref_text na „počet" v *a.*} <u>*celků*</u> {coref_text na „celek" v *b.*} *je zřejmě fetišem, který jsou koaliční lídři ochotni vyměnit za tu či onu pravomoc budoucích územně správních útvarů.*

d. Iracionalita politického procesu však akcentuje právě význam <u>*počtu*</u> {coref_text na „počet" v *c.*} <u>*celků*</u> {coref_text na „celek" v *c.*} *a umístění jejich center.*

e. Jinou otázkou zůstává skutečnost, že <u>*počet*</u> {coref_text na „počet" v *d.*} <u>*celků*</u> {coref_text na „celek" v *d.*} *a jejich kompetence se projednávají odděleně.*

f. Z tohoto pohledu byl již samotný návrh zásad ve středu schváleného zákona do značné míry paradoxní: pravomoci vytýčil <u>*celkům*</u> {coref_text na „celek" v *d.*}, <u>*jejichž*</u> {coref_gram na „celek" ve *f.*} <u>*počet*</u> {coref_text na „počet" v *e.*} *je dodnes ve hvězdách (vždyť až po schválení těchto zásad se např. objevil šokující návrh ODS na vytvoření osmdesáti regionů).*

Schématicky je naše rozhodnutí o anotování koreference v (112)a–f zobrazeno na obrázku č. 20:



Obrázek č. 20: Koreference u konstrukcí s kontejnerem

Podobná situace nastává v (113)a–e, kde obě jmenné fráze *počet* a *registrace* v (113)d referují odděleně:

(113) *a. Počet registrací nových osobních automobilů v Německu se v červenci proti předchozímu měsíci snížil o 16.3 procenta a v meziročním srovnání byl nižší o 4.9 procenta.*

b. Celkový počet {coref_text na „počet" v *a.*} *registrací* {coref_text na „registrace" v *a.*} *všech nových vozidel v tomto měsíci činil 299147.*

c. V červnu se počet {coref_text na „počet" v *b.*} *registrací* {coref_text na „registrace" v *b.*} *osobních automobilů snížil měsíčně o jedno procento a meziročně o 5.5 procenta.*

d. Za prvních sedm měsíců tohoto roku dosáhly registrace {coref_text na „registrace" v *c.*} *nových osobních vozů počtu* {coref_text na „počet" v *c.*} *2017474, což bylo o 0.1 procenta více než ve stejném období loni.*

e. Počet registrací počet {coref_text na „počet" v *d.*} *registrací* {coref_text na „registrace" v *d.*} *všech vozidel se však snížil o 0.3 procenta na 2355628 kusů.*

Srov. také anotaci dvou koreferenčních řetězců – se specifickou referencí u „kontejnerů" (*krabice – krabice*) a generickou u závislých členů (*guma – guma*) v (114)a–d:

(114) *a. V rámci reklamní kampaně předal osobně každému členu washingtonského Kongresu jednu krabici* {specifická reference, distributivní} *své gumy* {generická reference} *značky Yucatan.*

b. I král dostal svou krabici {specifická reference, žádná šipka} *gumy* {generická reference, coref_text, typ=NR na „guma" v *a.*} *a nádavkem i doslova trhoveckou prezentaci.*

c. Samozřejmě, že novinové zprávy o králi s krabicí {specifická reference, coref_text, typ=0 na „krabice" v *b.*} *gumy* {generická reference, coref_text, typ=NR na „guma" v *b.*} *byly reklamou k nezaplacení.*

d. I když konzervativní Anglie jeho čin odsoudila, guma {generická reference, coref_text, typ=NR na „guma" v *c.*} *se zde chytila a Británie se pro žvýkačku stala bránou do Evropy.*

Strategie řešení jednotlivých případů dovoluje formulovat následující **pravidlo anotace textové koreference v konstrukcích s „kontejnery“**:

Při výběru mezi anotací textové koreference na kontejner nebo na jeho závislý uzel, koreferenční vztah vedeme na kontejner. Pokud se však na kontejneru závislý uzel zúčastní koreferenčního vztahu samostatně, vyznačíme oba vztahy – mezi kontejnery a mezi závislými uzly.

III.4.2.3.4. Dvě místní určení vedle sebe (*tady v Praze, u nich doma* apod.)

Pokud v tektogramatickém stromě jsou dvě místní určení jako sestry a nejsou přitom ve vztahu apozice ani koordinace, propojíme je textovou koreferencí postupně po sobě. Srov. (115):

(115) *a. Na stůl přinášel kuchyni studenou , chlebíčky , uzeniny , šunku a západoněmecké sýry mnoha druhů a zde jsem žasl nad jejich kvalitou, kterou Petr z domova nepředpokládal.*
b. Při tlumeném světle přicházela na přetřes politická situace u nich[106] {coref_text, typ=0 na „domov“ v *a.*} *doma* {coref_text, typ=0 na `#PersPron` (*u nich*)}.

Pokládáme však za zbytečné propojovat textovou koreferencí upřesňující časová určení, protože se v tomto případě často jedná o komplikovanější vztah, ne vždy koreferenční. Srov. např. v (116) je problematické označovat koreferenci mezi *dnes* a *16 hodin*. Přesně vzato, tyto výrazy nejsou koreferenční.

(116) *Třídenní koncert nazvaný Trutnov 87 – 94 začíná dnes v 16 hodin* {žádný koreferenční vztah} *v trutnovském letním kině Na bojišti.*

III.4.2.3.5. Technické problémy

V následující kapitole se zaměříme na tři nesrovnalosti anotace koreference a syntaktické struktury tektogramatického stromu. V každém z těchto případů půjde o technickou nemožnost přesně zachytit koreferenční vztah bez překročení rámců tektogramatické roviny.

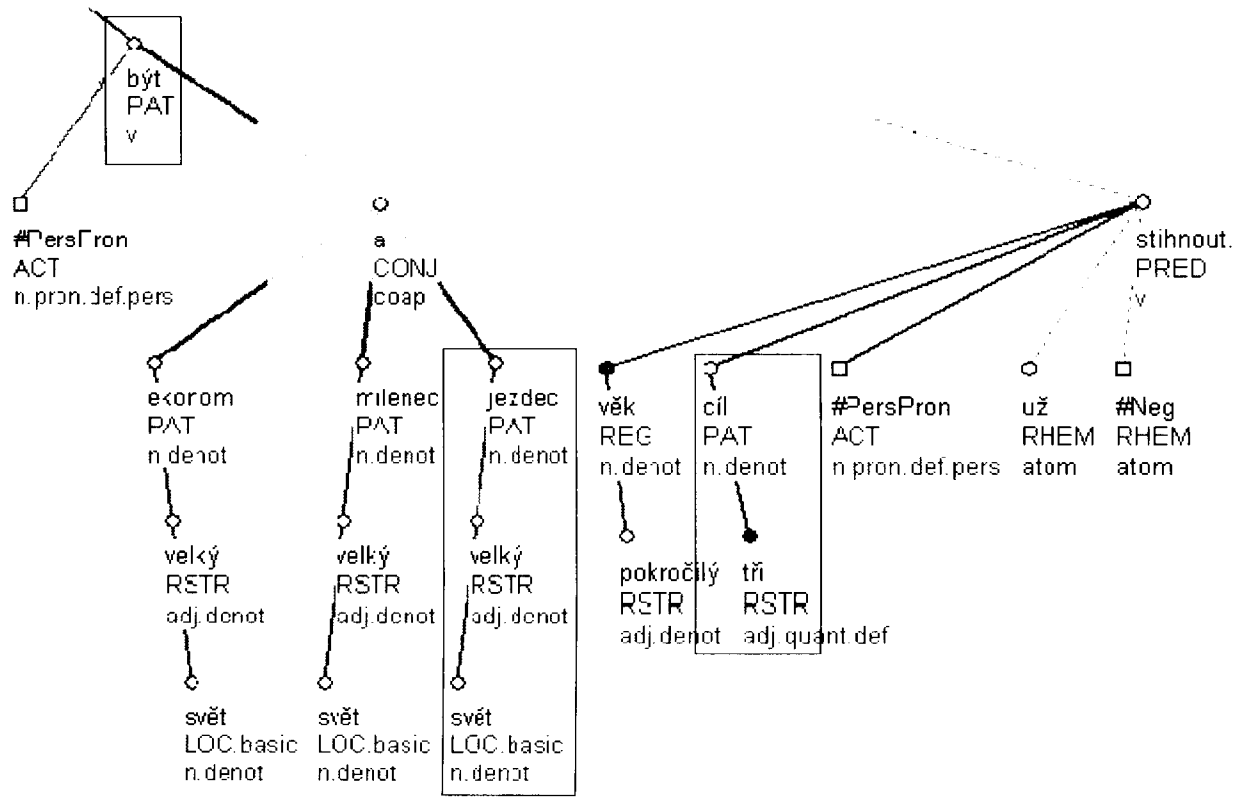
[106] Osobní zájmeno *jich* je v tektogramatickém stromě reprezentováno jako `#PersPron`.

III.4.2.3.5.1. Odkaz na nevyčlenitelný podstrom

V některých případech potřebujeme odkázat na úsek textu, který netvoří na tektogramatické rovině vyčlenitelný podstrom. Srov. (117)a–b, kde antecedentem NP *třetí cíl* je slovesná fráze *být největším jezdcem na světě.*

(117) *a. Na Harvardu , kde tento třešťský rodák učil od raných třicátých let až do své smrti, kolovala autobiografická poznámka: Chtěl jsem být největším ekonomem na světě, největším milencem na světě a největším jezdcem na světě.*
b. Vzhledem k pokročilému věku třetí cíl {coref_special, typ=segm} *už nestihnu.*

V tektogramatickém stromě však není takový podstrom – uzel *být* je použit jenom jednou pro všechny členy koordinační konstrukce (srov. obrázek č. 21).



Obrázek č. 21: Odkaz na neoddělitelný podstrom

Existují čtyři možná řešení:

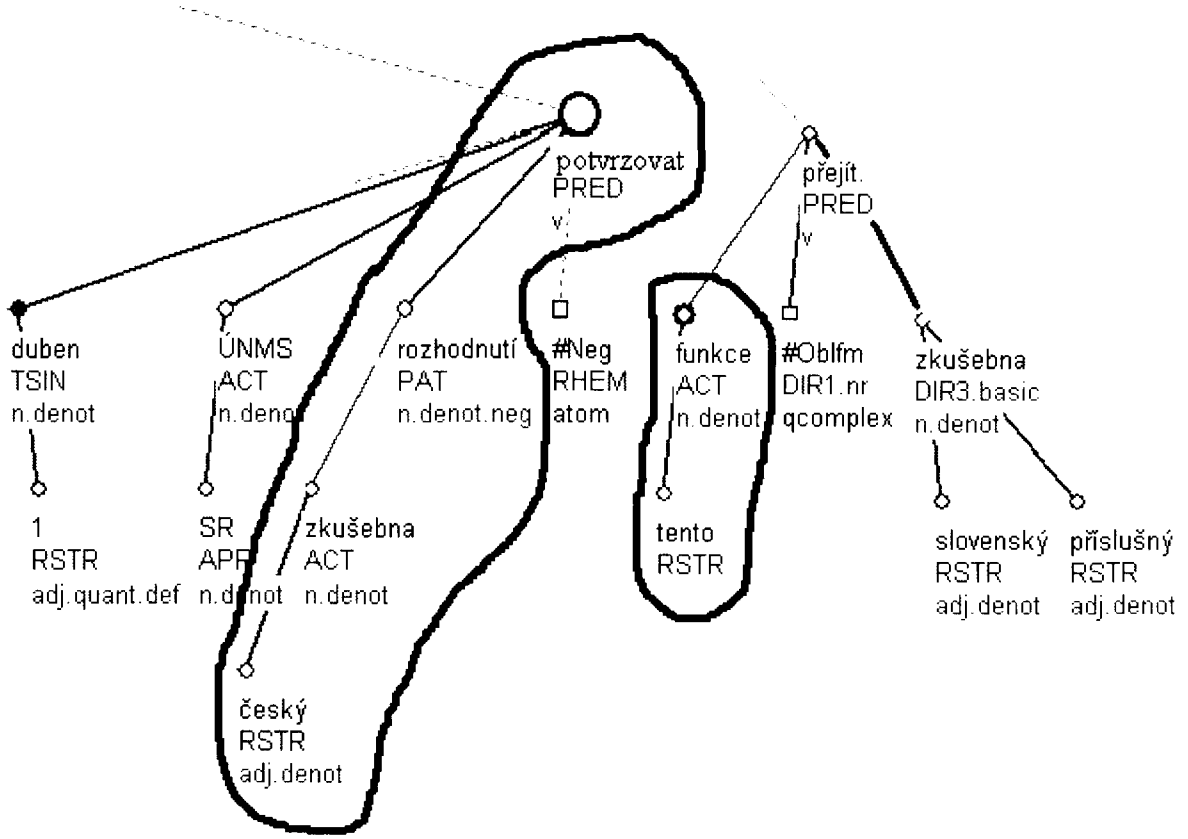
a) odkázat na celý podstrom s kořenem *být*. Tím se do podstromu antecedentu dostanou také jiné dva cíle, na které se neodkazuje v anaforické NP – *být největším ekonomem na světě* a *být největším milencem na světě*;

b) odkázat pouze na jmennou frázi *největší jezdec na světě.* Tím se z podstromu antecedentu ztratí slovesná složka a koreferující pár bude vypadat jako *největší jezdec na světě – třetí cíl*;
c) odkázat přesně na antecedentní slovesnou frázi. Toto sice není možné udělat v rámci anotace na tektogramatické rovině, ale mimo ni je to technicky proveditelně;
d) nehledat přesný antecedent a odkázat dozadu speciální koreferencí typu `segm`.

Je zřejmé, že první dvě řešení podstatně zkreslují skutečnost – ani v jednom případě nejde o koreferenci v tom smyslu, jak jsme ji definovali v předchozích kapitolách. Varianta c) by byla samozřejmě přijatelná, ale vyžaduje podstatnou změnu teoretické báze anotace. De facto, rozhodneme-li se pro tuto variantu, budeme muset řešit koreferenci závislých členů uzlů anaforu a antecedentu i ve všech ostatních případech, čili u každého koreferujícího uzlu je nutné se zamyslit nad tím, jestli všechny členy, které jsou na něm závislé, patří do koreferenčního vztahu. Taková práce pravděpodobně výrazně upřesní sémantickou stránku anotovaných koreferenčních vztahů. Na druhé straně realizace tohoto úkolu je mimořádně časově náročná. S největší pravděpodobností výrazně sníží mezianotátorskou shodu a tedy i užitečnost celé anotace pro aplikační úkoly. Kromě toho, tak složitou a detailní práci prozatím není možné provést automaticky. Vybíráme tedy variantu d). Jde o prozatímní řešení. V případě nutnosti, se tyto případy dají vyhledat podle relativně jednoduchých kritérií a zpracovat jiným způsobem.

Podobně v (118)a–b – struktura stromu v (118)a nedovolí odkázat NP *tato funkce* v (118)b na antecedentní slovesnou frázi *potvrzovat rozhodnutí české zkušebny*, pokud zůstaneme v rámcích tektogramatické roviny. Podobně jako v předchozím příkladě (117), řešíme to v anotaci zatím jako „coref_special, typ=`segm`". Srov. (118)a–b a obrázek č. 22:

(118) *a. Od 1. dubna nebude ÚNMS SR rozhodnutí české zkušebny potvrzovat.*
b. Tato funkce {coref_special, typ=`segm`} *přejde na příslušnou slovenskou zkušebnu, která bude vydávat na základě dodaných podkladů příslušné certifikáty.*



Obrázek č. 22: Odkaz na neoddělitelný podstrom

Srov. také analogickou situaci v (119)a–b:

(119) *a. Šedesát tři vězňů, kteří vykonávají trest odnětí svobody v České republice, požádalo za první půlrok o předání do věznic na území Slovenska.*
b. Loni podalo tuto žádost {coref_special, typ=segm} *200 odsouzených.*

III.4.2.3.5.2. Různé aspekty koreference – PP nebo NP

Samostatným problémem je anotace koreference u výrazů s předložkou. Předložky nejsou v tektogramatickém stromě reprezentovány uzlem. Jejich význam se na tektogramatické rovině promítá do hodnoty funktorů a subfunktorů (viz Mikulová a kol. 2005, s. 914). Nicméně, velká část jmenných frází s předložkou má jinou referenční hodnotu, než stejné jmenné fráze bez předložky. Srov. např. stejný referenční potenciál frází v (120) a různý v (121):

(120) *v Praze – o Praze – pro Prahu – bez Prahy apod.*

(121) *v Praze – před Prahou – za Prahou – pod Prahou – nad Prahou* apod.

Problém spočívá v tom, jak anotovat/neanotovat koreferenci mezi NP a PP v (121). Tento problém nevzniká při anotaci koreference na složkové struktuře. Jmenná fráze bude vnořena do předložkové fráze a koreferenčním vztahem se může propojit jak předložková, tak vnořená jmenná fráze. Také v případě závislé tektogramatické struktury, kde sémanticky nevyprázdněné předložky jsou reprezentovány samostatnými uzly, takový problém nevznikne (srov. Melčuk 1974).

Absence vyjádřených předložek na tektogramatické rovině nás implicitně vede k anotaci koreference u výrazů bez ohledu na předložky, se kterými jsou použity, tj. v (120) a (121) propojujeme všechny výrazy identickou textovou koreferencí. Tím bychom dodrželi relativní jednotnost anotace na úkor vyznačení skutečné koreference. Jiným řešením je orientace na reference předložkových frází jako celku, to je však výrazně časově náročnější a s největší pravděpodobností bude příčinou většího počtu chyb (anotátor si nevšimne předložky a propojí Prahu s Prahou, i když v jednom případě bylo *za Prahou* a v druhém – v *Praze*) a je těžko automaticky zachytitelné. Předpokládáme však, že alespoň částečně je možné referenční odlišnosti předložkových frází „vytáhnout" z tektogramatického stromu automaticky na základě významů subfunktorů koreferenčních uzlů[107].

Konvence o anotaci koreference předložkových frází tedy zní:

Pokud dvě jmenné fráze v textu jsou koreferenční, anotujeme jejich vztah jako textovou koreferenci, i v případě, že jsou součástí předložkových frází, které mezi sebou koreferenční nejsou.

Rozebereme příklady (122)–(124).

V (122)a–c je řetězec *za Prahu – části města – tu – z Prahy*. *Město* v (122)a je koreferenční s *Praha* v (122)a, *tu* však neznamená *v Praze*, nýbrž *za Prahou*. Obě anaforické fráze – *město* v (122)a a *tu* v (122)b – odkazují na tektogramatický uzel *Praha*, avšak *město* je koreferenční s Prahou, zatímco *tu* je koreferenční s územím za Prahou. Technicky však takové dvojí pojetí antecedentu nelze zachytit. Naše pojetí koreference jako symetrického vztahu zobrazeného řetězcem a automatické dodržování koreferenčních řetězců vypracované na základě tohoto pojetí způsobí, že pokud odkážeme *tu* pouze na *za Prahou*, automaticky se spojí s *město*.

(122) *a. Zatím se posunuje stále více za Prahu, čímž ztrácí na své účelnosti z hlediska dopravních spojení do jednotlivých částí města* {coref_text, typ=0 na „Praha"}.

[107] O subfunktorech viz Mikulová a kol. 2005, s. 591n.

b. Na druhé straně by tu {coref_text, typ=0 na „město" v *a.*} *asi mohlo být víc pozemků vhodných k podnikání.*

c. Po dálnici bychom se měli svézt z Prahy {coref_text, typ=0 na „tu" v *b.*}*až do Českých Budějovic, v roce 1997 pravděpodobně projedou první vozidla po dálnici Praha – Plzeň, dokončena by měla být i dálnice D8 z Prahy do Ústí nad Labem.*

Kromě místních určení, se tento problém týká určení časových. Např. výrazy *před druhou světovou válkou* a *po válce* v (123)a–b. Oba tyto výrazy jsou reprezentovány uzly s t-lemmatem *válka*, rozdíl mezi nimi je zachycen pouze subfunktorem (`before/after`)[108]. Je tedy otázkou, jak anotovat vzájemný vztah. Pokud bychom vycházeli z lemmatu toho uzlu, charakterizovali bychom jej jako koreferenci; pokud bychom zohlednili i informace ze subfunktoru, přiklonili bychom se spíše k asociační anafoře. V tom případě by se už kvůli dodržování konzistence anotace nemohla značit ani koreference mezi uzly *před válkou* a *válka* (např. ve větě, kde by *válka* byla podmětem), čímž by byla anotace výrazně ochuzena. Proto se podobně jako v předchozích příkladech řídíme i zde stanoveným pravidlem a anotujeme identickou textovou koreferenci mezi řídícími uzly *válka – válka.*

(123) *a. Před druhou světovou válkou pracoval pro československou armádu na zdokonalení zaměřovačů protiletecké obrany.*

b. Po válce {coref_text, typ=0 na „válka" v *a.*}*se Svoboda vrátil na ČVUT, kde kolem sebe shromáždil studenty a mladé vědecké pracovníky – první generaci našeho počítačového výzkumu.*

Dalším zajímavým příkladem je dvojice *před začátkem regionálního utkání* a *při rozcvičování* v (124)a–b. Zde je situace obrácená – pokud vezmeme v úvahu i význam předložky, zachycený subfunktorem, je to koreference, nicméně mezi referenty samotných substantiv žádný označitelný vztah není, a proto jej neanotujeme.

(124) *a. Krátce před začátkem regionálního utkání v portugalském Pico de Regalados zemřel 31letý brankář Albino Silva Ribeiro.*

b. Při rozcvičování {žádný koreferenční vztah} *chytil několik míčů a poté se v brance zhroutil.*

[108] Ibid. s. 603–604.

Typologicky podobné mohou být i koreferenční NP s místním významem – např. *před domem – na zahradě*, které nejsou daleko od párů s odkazovacími výrazy typu *tady – v Praze* a *u nich doma,* kde případnou koreferenci už běžně anotujeme.

III.4.2.4. Nejednoznačný výběr antecedentu

Pro zachování systematické důslednosti anotace je nutné přijmout několik konvencí o určování antecedentu v případech, kde se intuitivně nabízí více možností. V následující kapitole stanovíme pravidla pro určování antecedentu při odkazování z/na apoziční a koordinační skupiny (III.4.2.4.1.– III.4.2.4.2.). V III.4.2.4.3. rozebereme jeden typický příklad, kde je více možných antecedentů a rozhodnutí závisí na hlubší sémantické a referenční analýze textu.

III.4.2.4.1. K otázce výběru antecedentu v případe apoziční skupiny

Pro zachování jednotnosti anotace původní pronominální a rozšířené textové koreference a na základě principu o maximální velikosti koreferujících členů (III.1.3.) při anotaci koreference v konstrukcích s apozicí se řídíme následujícím pravidlem:

Pravidlo o výběru antecedentu v případe apoziční skupiny:

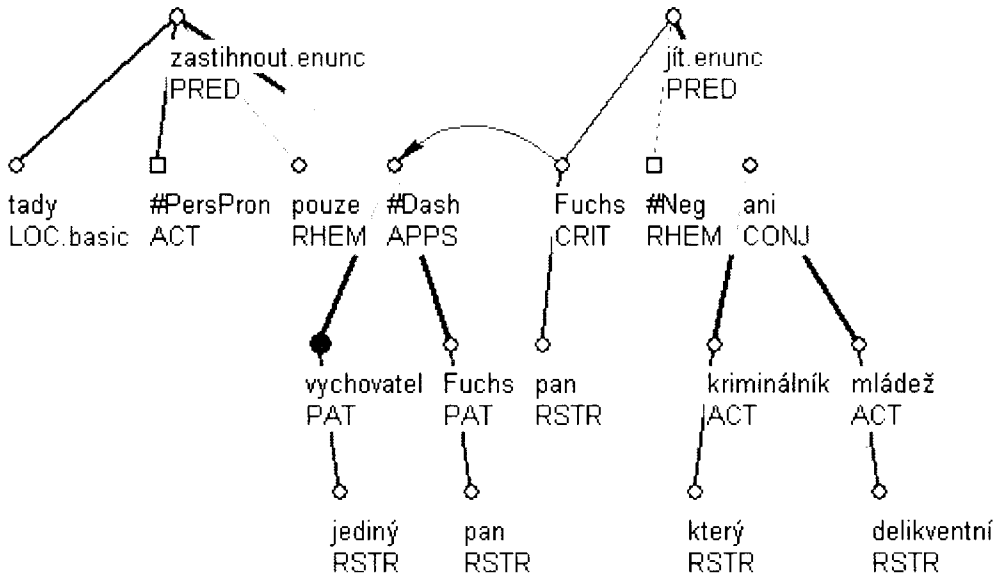
Pokud antecedent (anafor) je přímý potomek uzlu s funktorem `APPS`, koreferenční šipka vede na uzel s funktorem `APPS` (resp. od něj).

Takové řešení se může zdát arbitrární a poněkud nepřehledné – do koreferenčních řetězců se tak dostává velký počet spojek a rekonstruovaných uzlů pro interpunkční znaménka. Je to však jediný způsob jak zachovat jednotnost anotace.

Srov. následující příklady:

- koreference *NP – apoziční konstrukce*:

(125) *a. Zastihli jsme tady pouze jediného vychovatele – pana Fuchse.*
b. Podle pana Fuchse {coref_text, typ=0 na pomlčku (uzel s t_lemmatem `#Dash`), nikoliv na „pan Fuchse"} *nejde o žádné kriminálníky ani delikventní mládež.*



Obrázek č. 23: Koreference s apoziční konstrukcí

(126) *a. Tato slova řekl Raimund Strathman, muž z Evangelické vesnice mládeže, která má sídlo v Rensburgu v německé spolkové zemi Šlesvicko – Holštýnsko.*
b. Jsme domov mládeže, váš 'děcák', nikoliv 'pasťák', říká R. Strathman {coref_text, typ=0 na #Comma, nikoliv na „Strathman"}.

- koreference *apoziční konstrukce – NP*:

(127) *a. „Můj Williams-Renault prokázal technické nedostatky, reagoval nervózně a Schumacherovu Benettonu jsem mohl jen závidět," posteskl si v článku Senna a zamyslel se nad dravým mládím za volantem: „Barrichellova nehoda a smrt Ratzenbergera mě utvrdily, že s přibývajícími nováčky stoupá nebezpečí."*
b. Sennovy problémy postřehl v závodě i Michael Schumacher, {coref_text, typ=0 od #Comma na „Schumacher" v *a.*} *pozdější vítěz VC: „Jezdil jsem za Ayrtonem a v šestém kole jsem si všiml, že jeho vůz v té zatáčce maličko ztrácel stabilitu, o kolo později se tam vyboural."*

III.4.2.4.2. K otázce výběru antecedentu v případě koordinační skupiny

V případě identické textové koreference s koordinační konstrukcí, podobně jako v případě apozice (III.4.2.4.1.) preferujeme odkaz na formálně řídící uzel s funktorem CONJ. Je to podmíněno dvěma příčinami: a) zachováním jednotnosti anotace původní pronominální a rozšířené textové koreference a b) principem o preferenci identické koreference před asociační anaforou.

Můžeme tedy formulovat následující **pravidlo o výběru antecedentu při koreferenci s koordinační konstrukcí**:

Pokud anaforická NP je koreferenční s více elementy, které jsou syntakticky spojeny do koordinační konstrukce, navazujeme vztah identické textové koreference (nikoliv však asociační anafory) mezi anaforickou NP a formálním řídícím uzlem dané koordinační konstrukce s funktorem CONJ.
Podobně postupujeme při opačném pořadí členů.

Srov. následující příklady:

- koreference *koordinační konstrukce – NP*:

(128) *a. Nedělní vystoupení José Carrerase a Montserrat Caballéové v pražském Rudolfinu (na rozdíl od večera Plácida Dominga ve Sportovní hale) tuto potřebu skutečně naplnilo.*
b. Hudba zněla živě, bez elektroakustického zprostředkování a velikost sálu zároveň poskytovala možnost uspokojivého zrakového kontaktu s umělci {coref_text, typ=0 na spojku „a" v *a.*}.

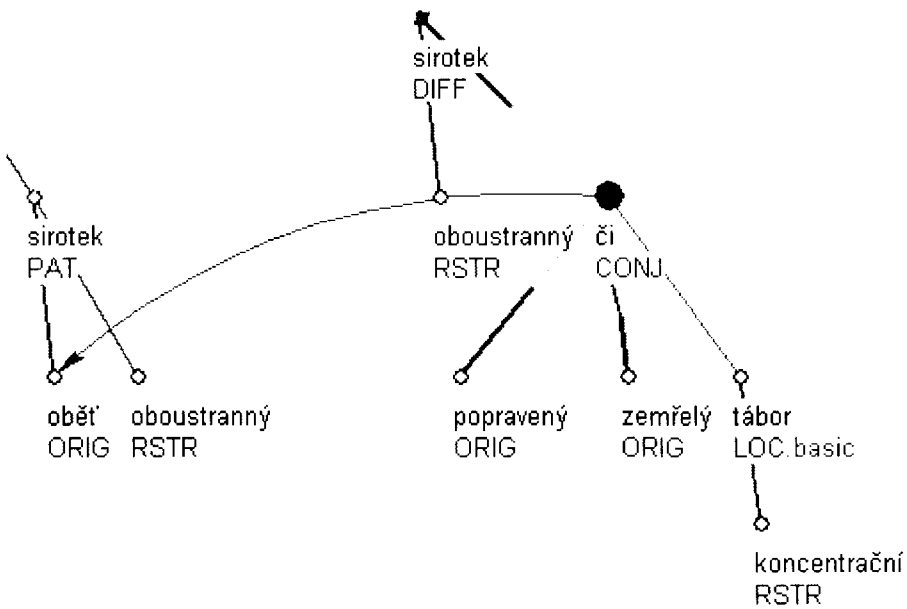
- analogický případ koreference *koordinační konstrukce – NP* v anotaci původní zájmenné koreference:

(129) *a. Vznik moderního umění se spojuje s rokem 1907, kdy byla založena populární Osma, a v kubistickém a fóbistickém duchu malují Filla, Kubišta, Špála a další.*
b. Ceny jejich {coref_text, typ=0 na spojku „a" v *a.*} *obrazů šplhají do statisíců a dobře se prodávají v cizině .*

- koreference *NP – koordinační konstrukce*:

(130) *a. Odškodněni by měli být i oboustranní sirotci po obětech.*

b. O oboustranné sirotky po popravených či {coref_text, typ=0 od spojky „či" na „oběť" v *a.*} *zemřelých v koncentračních táborech by se měl rozšířit okruh osob, jimž má být poskytnuta jednorázová finanční částka za perzekuci v době druhé světové války.*



Obrázek č. 24: Koreference koordinačních konstrukcí

- kataforické odkazování NP – koordinační konstrukce:

(131) *Z 1. pol. 19. století, kdy se rodila moderní česká krajinomalba, je dnes zájem hlavně o tyto autory: Otec a syn Mánesové, Navrátil, Piepenhagen, Kosárek, Bubák, Ullík, Havránek.* {coref_text, typ=0 od čárky `#Comma` na „autor"}

III.4.2.4.3. K otázce více možností odkazování (identická koreference)

Ne vždy se setkáme s příklady, které se dají snadno klasifikovat a tedy důsledně zařadit. V následujícím příkladě (132)a–g mají jmenné fráze s generickou referencí více než jeden možný antecedent. Rozhodli jsme se pro dva generické koreferenční řetězce, avšak toto rozhodnutí je

založeno téměř výhradně na intuici a je velice diskutabilní. Koreferenční ambiguitu komplikuje také skutečnost, že se koreferenčního vztahu zúčastní výrazy označující hmotu měnící se v čase.

(132) *a. Pro historii žvýkací gumy, jak ji známe dnes, se však musíme přenést na jiný kontinent.*

[...: výklad o tom, jak Mayové vyráběli žvýkačku...]

b. Mízu stromu sapodilla (achras sapota) sklízeli a upravovali systémem, který se používá dodnes.

c. Kůru stromu nařízli do tvaru písmene v a do špičky řezu umístili nádobu, do níž šťáva {coref_text, typ=NR na „míza" v *b.*} *ukapávala.*

d. Získanou mléčnou gumovitou látku {coref_text, typ=NR na „šťáva" v *c.*} *pak čistili, vařili.*

e. Teprve výsledný substrát {coref_text, typ=NR na „látka" v *d.*} *byl hoden žvýkání.*

f. Zrodila se žvýkačka {coref_text, typ=NR na „guma" nebo coref_text, typ=NR na „substrát"}.

g. Tak jako každý Mexičan, i Santa Anna znal a občas žvýkal mízu {coref_text, typ=NR na „žvýkačka" v *f.* nebo „substrát" v *e.*} *sapodilly zvanou chicle ....*

Zdá se logické propojit jak *substrát* v (132)e a *žvýkačka* v (132)f, tak i *žvýkací guma* v (132)a a *žvýkačka* v (132)f, protože větou (132)f se končí výklad o její výrobě. Tím by však vznikly dva řetězce textové koreference, které by se křížily, což nelze ani technicky ani logicky (vycházíme z předpokladu, že koreferenční vztah je symetrický a tranzitivní, viz III.4.). V daném případě by to odpovídalo kohezi textu. V druhé posloupnosti (*míza – šťáva – látka ... míza*) máme koreferenční řetězec *míza* → *šťáva* → *látka* ve větách (132)b–d, potom následuje *míza* v (132)g, která odkazuje k *míza* v (132)b a tedy podle pravidel textové koreference je koreferenční se *šťáva* a *látka*. V daném kontextu však odkazuje na *míza* v (132)b, nikoliv na *látka* v (132).

Podobné případy nelze důsledně klasifikovat, proto postupujeme vždy podle intuice.

III.5. Asociační anafora (bridging vztah)

Asociační anaforu (bridging anaphora, indirect anaphora, bridging vztah) chápeme jako určitý nekoreferenční sémantický nebo pragmatický vztah mezi blízkými výrazy v textu, podílející na koherenci daného textu (k vysvětlení podstaty daného pojmu viz III.1.).

Termín *bridging* byl navržen v článku H. H. Clarka (Clark 1977), jako navazování spojovacího „můstku" k předchozím výrazům. V kognitivní lingvistice se používá termín *indirect anaphora* kvůli vyhnutí se významu vyvozování jednoho významu na základě jiného. V české tradici daný jev nebyl zkoumán systematicky, tedy neexistuje jeden vžitý termín, který bychom mohly přímo přijmout. V naší práci jako prozatímní řešení používáme termín *asociační anafora.* Jsme si vědomi toho, že popisovaný vztah mezi členy páru ve většině případů není anaforický a ani není založen na asociaci. Nicméně se nám nepodařilo najít vhodný překlad pro termín *bridging* Clarka, který pokládáme za nejvýstižnější pro popis daného jevu. V našich výkladech rovněž občas používáme přímo anglický termín *bridging* ve spojení s českým *vztah*, který podle našeho názoru je v daném případě vhodnější než *anafora.*

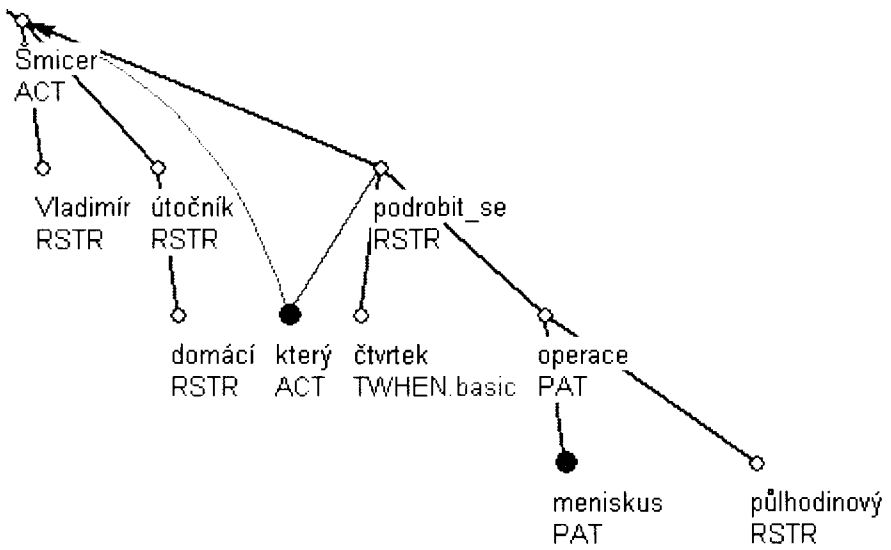
Při anotaci asociační anafory odkazujeme anaforický člen na nejbližší uzel antecedentního koreferenčního řetězce. Vycházíme totiž z předpokladu, že všechny elementy uvnitř koreferenčního řetězce jsou rovnocenné, není tedy důležité, na který uzel povede vztah asociační anafory. Vztahy asociační anafory jsou definovány na základě jejich referenčního potenciálu, nikoliv na základě pojmenování v konkrétním jazykovém projevu, což také odůvodňuje naše rozhodnutí. Na povrchové rovině však situace anotace asociační anafory vypadá jinak: tento vztah je výrazně méně formální než textová koreference. Určování asociační anafory vyžaduje veškeré znalosti mimojazykové skutečnosti, sémantiky výrazu a jejich konceptuálních, hypo- a hyperonymických, meronymických a jiných vztahů v jazykovém systému a také znalost (občas i pragmatického) kontextu. Jsou to informace, které se týkají především autosémantických výrazů, a proto i samotný vztah se de facto vyznačuje mezi nimi. Následně probíhá automatická kontrola zaznamenaných vztahů a upravuje se tak, aby vztah asociační anafory vedl k nejbližšímu anaforu uzlu koreferenčního řetězce antecedentu (viz IV.1.2.2.). Pokud vztahy mezi autosémantickými výrazy budeme pro případnou analýzu potřebovat, lze je bez problémů získat.

Formální povrchovou realizaci páru ve vztahu asociační anafory lze shrnout do následujícího schématu:

1. Anafor asociační anafory může být:
   a) lexikálně vyjádřený autosémantický uzel (*člověk, podmínka, skutečnost* atd.);
   b) seskupení lexikálně vyjádřených autosémantických uzlů spojených do koordinační nebo apoziční konstrukce, tedy formálně spojka nebo interpunkční známenko (*pokoj – strop, podlaha a zdi*);
   c) ostatní – všechno, co může být antecedentem (viz 2.).
2. Formálním antecedentem asociační anafory je poslední (nejbližší k anaforu) uzel identického koreferenčního řetězce (včetně uzlů, připojených gramatickou koreferencí) skutečného antecedentu vztahu asociační anafory.

Srov. např. (1), kde vztah část – celek mezi *Šmïcer* a *meniskus* je „roztržen" vztažným zájmenem *který*, které je spojeno s NP *Šmicer* gramatickou koreferencí:

(1) *Šestigólovou výhru Slavie nad Brnem ve 24. ligovém kole sledoval z tribuny i útočník domácích Vladimír Šmicer, který* {coref_gram na "Šmicer"} *se ve čtvrtek podrobil půlhodinové operaci menisku* {bridging WHOLE_PART na „který"}.



Obrázek č. 25: Asociační anafora – odkazování na poslední uzel koreferenčního řetězce antecedentu

Pro určování asociační anafory nehraje roli forma označujících členů páru. Elementy, mezi kterými postulujeme vztah, mohou být lexikálně stejně vyjádřeny, ale přitom nebýt

koreferenční (srov. (2)a–b – (3)a–b). A naopak povrchová textová realizace párů může navádět na asociační anaforu, zatímco ve skutečnosti jde o identickou (textovou) koreferenci (Srov např. (4) a–b).

(2) *a. Doufejme, že linka si časem vydobyde mezi dětmi takovou autoritu, aby se na ni obracely i ty, které jsou skutečně ohrožovány.*
b. Když si dítě {bridging SET_SUB[109] na „dítě" v *a.*} *bude přát, aby se o jeho problému nikdo z rodiny nebo školy nedozvěděl, musíme to respektovat, vysvětluje Jana Drtilová.*

(3) *a. Určitou svou představu si chci ověřit a potvrdit při cestě po USA, kterou jsem obdržel za vítězství v soutěži podnikatelů* {„soutěž" má specifickou referenci}.
b. K různým soutěžím {bridging SUB_SET na „soutěž" v *a.*} *mám výhrady.*

(4) *a. Péťa skončil u Jany Drtilové.*
b. Všechno nakonec dobře dopadlo, ale tohle dítě {coref_text, typ=0 na „Péťa" v *a.*} *zbytečně prožilo půl roku strachu a děsivých představ.*

Vztahem asociační anafory odkazujeme anaforicky na předchozí kontext. Toto rozhodnutí bývá často do jisté míry arbitrární, zakládá se na obecných znalostech o komunikativní organizaci textu a na principu jednotnosti anotace. Jsou však možné jednotlivé výjimky, kdy šipka asociační anafory vede na uzel v následujícím textu, a to v případě skutečného kataforického vztahu, na který poukazují explicitní odkazovací prostředky (především ukazovací zájmeno). Srov. následující příklad (5)a–e asociační katafory:

(5) *a. Ministr Karel Dyba: Vzhledem k očekávané poptávce a dané sumě, která je k dispozici ze státního rozpočtu , byly v nových programech Ministerstva hospodářství pro rok 1994 provedeny následující změny* {bridging, typ=SET_SUB,[110] na kořeny stromů (5)b–e}:
b. Za a). S výjimkou programu Region a Aeskulap byla snížena sazba příspěvku na úhradu úroků o 1–2% proti roku 1993.
c. Za b). Byla zkrácena doba od podpisu úvěrové smlouvy k registraci žádostí o

[109] K významům typů bridging vztahů viz III.5.1.1. – III.5.1.6.

[110] V případě bridging katafory pořadí vztahů v typech SET, PART a FUNCT označujeme podle lineárního pořadí výrazů v textu.

podporu z 1 roku na 6 měsíců.

d. Za c). Byly zrušeny cenově zvýhodněné záruky za úvěr u jednotlivých programů a vyčleněny do nového programu Záruka.

e. Za d). Byl zrušen příspěvek na úhradu úroků pro obchodní činnost s výjimkou programu Start.

III.5.1. Typologie vztahů asociační anafory

Naše klasifikace typů vztahů asociační anafory je vytvořena na základě existujících klasifikací, zaměřených na anotaci s podobným nebo stejným aplikačním účelem. K typologii vztahů asociační anafory v rámci kognitivní lingvistiky viz III.1. Typologie zaměřené na anotaci textů jsou představeny v klasickém článku Clarka (Clark 1977) a v současných klasifikacích, zaměřených na anotaci velkých textových korpusů v DRAMA (Passonneau 1996), Müller – Stube (2001), Gardent (2003), MATE/GNOME (Poesio 2004 aj.), PoCoS (Chiarchos – Krasavina 2005) aj.

H. Clark vyčleňuje následující nekoreferenční vztahy typu bridging:

- množina – podmnožina (*I met two people yesterday. The woman told me a story.*);
- nepřímý vztah na základě asociace, který má širokou škálu významů. Clark zmiňuje následující tři:

  i.neodlučitelné nutné části (*I looked into the room. The ceiling was very high.* – pokoj vždy má strop);

  i.možné části (*I went shopping yesterday. The walk did me good.* – *Go* ne vždy znamená *jít pěšky.*);

  i.části indukované na základě významu celku (*I went shopping yesterday. The climb did me good.*);
- nepřímý vztah na základě charakterizace. Jde především o vztah situace a její účastníků:

  ii.obligatorní aktanty situace (*John was murdered yesterday. The murderer got away*);

  iii.neobligatorní aktanty (*John died yesterday. The murderer got away*);

  iv.příčiny, výsledky, důvody apod.

Klasifikace Clarka je orientována na vysvětlení kontextové zapojenosti anaforických členů. Podobná, ale ještě detailnější klasifikace nekoreferenčních vztahů, kde kontextová zapojenost anaforického členu je podmíněna existencí v předchozím kontextu určitého antecedentu, je představena u Daneše (1999).

Klasifikace bridging vztahů podle Clarka a anaforických vztahů u Daneše jsou však zaměřeny na teoretický výzkum, jsou deskriptivní a přiblížené reálné situaci v jazyce. Proto typy vztahů uvedené v těchto pracích jsou spíše exemplifikační než vyčerpávající a jsou představeny jako rysy graduální povahy, které jsou navíc podrobně vnitřně diferencované. V naší klasifikaci jsme se na tyto dvě klasifikace orientovali, ale museli jsme je výrazně zjednodušit a formalizovat.

V anotačních schématech existujících v současné době jsou představeny následující klasifikace.

V projektu **DRAMA** R. Passonneau (1996) se anotuje velký počet vztahů asociační anafory, jako např. podmnožina, prvek množiny, část – celek, událost – příčina, objekt – vlastník, událost – argument apod., přičemž na typy výrazů, které se mohou zúčastnit těchto vztahů, nebyla kladena žádná omezení. Tak podrobná klasifikace se však prokázala jako vysoce subjektivní a nebyla aplikována na velká korpusová data.

Müller – Stube (2001) pro standardní anotační nástroj na koreferenci a asociační anaforu MMAX nabízí již pouze tři vztahy asociační anafory: událost – výsledek (*zkouška – výsledek*), část – celek (*dům – terasa*), a jednotka – atribut (např. v angličtině *John – 15 years old* (dosl. *John – patnáctiletý* v substantivní funkci)).

Gardent (2003) zakládá vztahy asociační anafory na sémantických vztazích. Nabízí následující systém typů:

- množina – podmnožina (*studenti – jeden student*),
- událost – argument (*vražda – vrah*),
- vyplývající ze slovníkové definice:
  - i.jednotka – atribut (např. angl. *John – 15 years old* apod.),
  - ii.meronymické vztahy (část – celek, celek – kus, událost – součást události apod.),
- vztah mezi účastníky jedné situace (angl. *jízda – sedadlo,* kde oba výrazy mají v lexikální sémantice komponent *„dopravní prostředek"),*
- nelexikální:
  - iii.okolností (místní nebo časové určení, např. *Grenoble – oblast*),
  - iv.vyplývající ze znalosti světa (*zima – sníh*).

V meta-schématu **MATE** s anotací asociační anafory se počítá v tzv. „třetím rozšíření" (Davies a kol. 1998, Poesio a kol. 2000d). Asociační anaforou se rozumí anaforické výrazy, které nebyly explicitně zmíněny v předchozím kontextu, ale které jsou v určitém sémantickém vztahu k některé entitě, která již v předchozím kontextu zmíněna byla, přičemž tento sémantický vztah není koreference. V tomto případě atribut TYPE odkazu <link> má jeden z následujících významů:

- Množina – prvek (Set – membership):

(6) angl. *The kids went to a party last weekend. Paul wanted to wear his new suit.*
č. *Děti byly minulý víkend na večírku. Paul si chtěl vzít nové sako.*

- Množina – podmnožina (Subset):

(7) angl. *You have here the models of rockets. And you are going to try to classify the rockets that have flown well.*
č. *Tady máte modely raket. Pokusíte se určit rakety, které letěly dobře.*

- Posese (Possession) s možnými podtypy: attribute, partitive, strict possession:

(8) angl. *The device, which included two booster tubes, may have been designed for an attack.*
č. *Toto zařízení, jehož součástí byly dvě posilovacítrubice, mohlo být navrženo pro útok.*

- Vázaná anafora (Bound anaphora):

(9) angl. *Nobody likes his job.*
č. *Nikdo nemá rád svoji práci.*

- Funkce – hodnota (Function – value):

(10) angl. *The temperature rose to 90 degrees.*
č. *Teplota stoupla na 90 stupňů*

- Konkretizace (Instantiation)

(11) angl. *We need oranges. There are some at Corning.*
č. *Potřebujeme pomeranče. Několik je Corningu.*

- Událost – sémanticky související entita (Event – relation):

(12) angl. *There was an explosion. The noise was tremendous.*
č. *Došlo k výbuchu. Hluk byl neuvěřitelný.*

Asociační anafory uvedených typů se ve schématu MATE anotovaly pouze v případě, že koreferenční antecedent nebyl nalezen. Avšak i při takových omezeních výsledky mezianotátorské shody byly relativně nízké (srov. Poesio – Vieira 1998).

Nejdůslednější aplikace meta-schématu MATE byla realizována v projektech GNOME (Poesio 2004a, 2004c a II.3.3.) a VENEX (Poesio a kol. 2008). V těchto korpusech se při anotaci asociační anafory vyčleňují následující vztahy:

1. ELEMENT (jmenná fráze referuje na prvek množiny vyjádřené antecedentem, např. *The sixteen panels are each divided into three horizontal zones, the middle containing a letter.*),
2. SUBSET (jmenná fráze referuje na podmnožinu množiny vyjádřené antecedentem),
3. POSS (posese v širokém smyslu; zahrnuje také vztah části a celku typu *dveře – byt*).

Aplikace anotace anaforických vztahů na korpusu GNOME byla evaluována ve dvou fázích. První evaluace (určovaní identické anafory, tj. koreferenčních výrazů se stejnou hlavou) byla provedena dvěma anotátory na 200 koreferenčních jmenných frázích a dosáhla poměrně výsoké mezianotátorské shody: 79.4% vztahů bylo označeno oběma anotátory, 12.8% vztahů bylo označeno jenom jedním anotátorem a v 7.7% případů jeden anotátor určil lineárně bližší antecedent než druhý anotátor. Avšak při anotaci asociační anafory stejným způsobem bylo označeno pouze 22% vztahů.

* * *

Představený rozbor odborných názorů na klasifikaci asociační anafory a analýza vztahů v konkrétních příkladech nás vedla k rozhodnutí o anotaci asociační anafory na tektogramatické rovině v PDT a k vyčleňování následujících šesti podtypů:

- meronymický vztah části a celku (PART_WHOLE a WHOLE_PART), viz III.5.1.1.;
- vztah mezi množinou a podmnožinou/prvkem množiny (SUB_SET a SET_SUB), viz III.5.1.2.;
- vztah mezi objektem a definovanou na něm unikátní funkcí (P_FUNCT a FUNCT_P), viz III.5.1.3.;
- vztah sémantického a pragmatického kontrastu (CONTRAST), viz III.5.1.4.;
- nekoreferenční anaforický vztah (ANAF), viz III.5.1.5.;
- blíže neupřesněná kategorie (REST), viz III.5.1.6.

V následujících oddílech III.5.1.1. – III.5.1.6. se chceme věnovat podrobnému popisu a příkladům vztahů těchto typů.

III.5.1.1. Vztah PART mezi částí a celkem (PART: PART_WHOLE a WHOLE_PART)

Meronymický vztah části a celku je jedním ze základních vztahů, který se vyčleňuje ve všech klasifikacích v dané oblasti. V naší anotaci jako PART anotujeme vztahy neodlučitelných částí k celku (resp. celku k částem).

Vztah PART je oboustranný, tj. vztah *celek – část* je stejně přínosný a důležitý pro koherenci textu jako *část – celek*. Sledujeme-li lineární pořadí výrazů v textu, existují následující možnosti:

- PART_WHOLE – první člen páru je část, druhý člen páru je celek;
- WHOLE_PART – první člen páru je celek, druhý člen páru je část.

Vzorové příklady jsou např. *pokoj – strop, ruka – prst, dům – pokoj, město – ulice* apod.

Srov. následující příklady (13)–(16) z PDT:

- vztah PART_WHOLE mezi *zatáčku Tamburello* a *okruh v Imole* v (13)a–b:

(13) *a. Trojnásobný mistr světa a autor tohoto výroku nepřežil VC San Marina, když v třísetkilometrové rychlosti nevytočil zatáčku Tamburello a narazil do betonové zdi.*
b. Po sobotní smrtelné nehodě Rakušana Rolanda Ratzenbergera si okruh v Imole {bridging, typ=PART_WHOLE, na „zatáčka“ v *a.*} *připsal druhý křížek a v Brazílii byl vyhlášen třídenní státní smutek.*

- vztah WHOLE_PART, např. vztah *budova – jednací síň, kluby, chodby* v (14)a–b:

(14) *a. [...] budova ČNR praskala ve švech, ovšem do roku 1992 to zajímalo jen málokoho.*

[... 5 vět ...]

b. Pokud tedy zrovna nesedí na svém minikřesle v jednací síni {bridging, typ=WHOLE_PART, na „budova" v *a.*}, *jsou poslanci nuceni pobývat buď ve svých klubech* {bridging, typ=WHOLE_PART, na „budova" v *a.*}, *nebo postávat či posedávat po chodbách* {bridging, typ=WHOLE_PART, na „budova" v *a.*}.

- oba směry PART_WHOLE a WHOLE_ PART mohou vytvářet jednu koherenční síť v jednom kontextu. Srov. v (15) vztah PART spojuje části patra paláce, v (15)a–b – Česko, Slovensko a Československo.

(15) *Kromě pracoven* {bridging, typ=PART_WHOLE, na „patro"} *bude v palácových patrech několik kuloárových chodeb, zasedacích místností a* {bridging, typ=WHOLE_PART od spoky „a" na „patro"} *přijímacích salonků s barokními stolky a křesly.*

(16) *a. Jejich vysílače dosud pokrývají signálem programu ČT 2 méně než polovinu území republiky* {jde o „ČR"}.

b. Do rozdělení federace {bridging, typ=PART_WHOLE, na „republika" v *a.*} *totiž signál zajišťovaly vysílače v SR* {bridging, typ=WHOLE_PART, na „federace"}.

Vztah PART anotujeme v následujících případech:

1. U jmenných frází, označujících území a části území[111] (typ *Německo – Bavorsko – Mnichov, město – ulice, Maroko – marocká města, Maroko – Marrákeš* apod. Srov. např. (16)a–b).
2. V prototypických případech neodlučitelných částí objektů, které nejsou myslitelné jako podmnožiny (*pokoj – strop, ruka – prst, pojišťovny – přepážka* apod., srov. (17)):

(17) *a. Jednotlivá studia v apartmánech jsou vybavena kuchyní* {bridging,

[111] V některých zdrojích je tento vztah vyčleňován zvlášť jako vztah Place/Area (srov. Gardent 2003).

typ=WHOLE_PART, na „studium"}, *takže je možná individuální příprava stravy.*

3. U odkazů na části časových úseků. Srov. (18):

(18) *a. Dělal jsem bez přestávky celé týdny, často v noci* {bridging, typ=WHOLE_PART, na „týden"}.

V ostatních případech asociační anaforu buď neanotujeme vůbec (typ *město – muzeum*) nebo anotujeme jako SUBSET (viz III.5.1.2.)

Jako **test na asociační anaforu typu PART** mezi objektem a částí objektu můžeme použít následující otázkový test:

U jmenné fráze X, u které předpokládáme, že je označovatelnou částí celku Y, se ptáme:

Je X ČÁST Y nebo je X NA (příp. V) Y?

Pokud odpovíme „je to část Y", jde o asociační anaforu typu PART.

Pokud odpovíme „Je to na (příp. v) Y", jde pravděpodobně o vztah, který nijak nezaznamenáváme.

Pomocí tohoto testu do PART zařadíme vztahy typu *stát – město; stát – region; město – ulice* apod., a vyřadíme vztahy typu *město – dům, město – muzeum*, které jako PART anotovat nechceme. Například vztah mezi městem a tím, co v tom městě je, jako asociační anaforu PART neanotujeme. Přesněji řečeno, muzeum není částí města a obraz není částí galerie. Srov. např. v (19), kde nebude označen žádný vztah typu PART:

(19) *a. V Mnichově jsou muzea* {žádný koreferenční vztah} *a galerie* {žádný koreferenční vztah} *se vzácnými obrazy* {žádný koreferenční vztah}, *částečně jsem je navštívil a zhlédl překrásný královský zámek Nymphenburg* {žádný koreferenční vztah}.

Podobně jako v (19), v následujícím příkladě (20)a–d stoly, pivo a jídelní lístek nejsou přesně řečeno součástí pivnice, ale jsou spíše *v* ní a těsně s ní souvisí. Tyto vztahy tedy neanotujeme.

(20) *a. Kdykoliv jsem navštívil Mnichov nebo jím projížděl, neopomněl jsem zajít do*

pověstné mnichovské pivnice Bierbräukeller .

b. Její obrovská hala, kde se sedělo u dubových stolů {vztah k „pivnice", žádný koreferenční vztah} *a bavorské pivo* {vztah k „pivnice", žádný koreferenční vztah} *se nalévalo do litrových korbelů s příklopkami.*

c. Jídelní list {vztah k „pivnice", žádný koreferenční vztah} *byl bohatý na zabíjačkové pochoutky* {vztah k „jídelní list", žádný koreferenční vztah} *a masa* {vztah k „jídelní list", žádný koreferenční vztah}, *dal se objednat talíř* {vztah k „jídelní list", žádný koreferenční vztah}, *kde bylo ode všeho něco, jitrnička* {vztah k „jídelní list", žádný koreferenční vztah}, *jelítko* {vztah k „jídelní list", žádný koreferenční vztah} , *klobása* {vztah k „jídelní list", žádný koreferenční vztah} *a plátek* {vztah k „jídelní list", žádný koreferenční vztah} *ovaru nebo větší porce* {vztah k „jídelní list", žádný koreferenční vztah} *zvlášť , ke všemu byl čerstvý rohlík* {vztah k „jídelní list", žádný koreferenční vztah} *nebo chléb* {vztah k „jídelní list", žádný koreferenční vztah} *a křen* {vztah k „jídelní list", žádný koreferenční vztah} *a hořčice* {vztah k „jídelní list", žádný koreferenční vztah}.

d. S pivem {vztah k „pivnice", žádný koreferenční vztah} *to bylo výtečné.*

V (21)–(22) však vztah PART mezi *Ovocný trh – Praha* a *Cunlhatu – střední Francie* ještě označíme. V obou případech jsou uzly v TGS sestry a tím jsou analogické se vztahem mezi dvěma místními určeními, které anotujeme v rámci identické textové koreference:

(21) *Dva víkendové dny plné zábavy, veselých písniček, soutěží a kouzlení pod názvem Vítejte z prázdnin připravila na Ovocném trhu v Praze* {bridging, typ=PART_WHOLE, na „trh"} *nadace Zdraví jako pozdrav pražským školákům.*

(22) *Sraz milovníků legendárních motocyklů Harley-Davidson v Cunlhatu ve střední Francii* {bridging, typ=PART_WHOLE, na „Cunlhat"} *měl podstatně vyšší účast než nedávné setkání majitelů strojů této značky v Praze.*

Dalším problematickým bodem v označování vztahu PART mezi územím a částí území je omezení hloubky vložení části do celku. Tak např. bychom chtěli označovat vztah PART v páru typu *kontinent – stát, stát – město, město – ulice, ulice – dům*, ale neanotovat např. *kontinent – město, stát – ulice, město – dům*. Nemůžeme však formulovat jednoduché pravidlo, že u asociační anafory typu PART může jít pouze o vložení prvního řádu, protože např. mezi

stát a město se může dostat další kategorie typu oblast, region, a jiné části větší než město (např. *Bavorsko* v *Německo – Bavorsko – Mnichov*), které bychom chtěli anotovat jako část (PART) státu a nezakázat přitom PART mezi státem a městem. Není proto možné stanovit omezující pravidlo v termínech kognitivní lingvistiky a zbývá pouze postupovat podle vlastní intuice a kontextu konkrétního anotovaného textu (tj. v případě, že se v textu jsou všechny tři kategorie – stát, region a město – anotovat město jako část regionu, nikoliv jako část státu).

III.5.1.1.1. Vztah PART uvnitř jedné věty

Vztah PART běžně vystupuje uvnitř jedné věty. Pokud meronymický vztah není již zachycen ve struktuře tektogramatického stromu (viz III.5.2.3.), zachycujeme ho šipkou asociační anafory. Srov. (23):

(23) *V ČR se objevily firmy, které dodávají vodoměry ve dvou částech – zvlášť počítadlo* {bridging, typ=WHOLE_PART, na „vodoměr"}, *zvlášť spodní část* {bridging, typ=WHOLE_PART, na „vodoměr"}, *která mívá poruchy.*

Vztah typu PART uvnitř jedné věty má převahu nad vztahem stejného typu s elementy jiné věty. Např. v (24)a–b označíme asociační anaforu typu WHOLE_PART mezi *suterény* a *budovy* v (24)b, nikoliv WHOLE_PART mezi *suterény* v (24)b a *nový komplex* v (24)a:

(24) *a. Žádné honosné taneční sály nebo restaurace v novém komplexu nebudou .*

[... několik vět ...]

b. Poslanci budou muset odpočívat jinde, protože suterény jsou příliš hluboko a napojení na výše položenou kanalizaci pomocí čerpadel by provoz budov {bridging, typ=WHOLE_PART na „suterén"} *neúměrně prodražilo.*

III.5.1.1.2. Vztah PART u generických NP a deverbativ

Asociační anafora typu PART je možná u také generických NP (srov. (25)a–b), je však v našem korpusu málo frekventovaná, občas splývá s asociační anaforou typu SUBSET a je vyřešena ve prospěch SUBSET.

(25) *a. Současný heroin je také mnohem čistší a jemnější než dříve.*

b. V běžném vzorku sedmdesátých let byla pouze 3–4 procenta čisté suroviny

{bridging, typ=WHOLE_PART na „vystoupení"}.

U deverbativ a abstraktnich výrazů jiného druhu je rozdíl mezi PART a SUBSET často eliminován a vyřešen v naší anotaci ve prospěch SUBSET (viz III.5.1.2.). Jsou však jednotlivé výjimky. Srov. např. (26)a–b:

(26) *a. Nedělní vystoupení José Carrerase a Montserrat Caballéové v pražském Rudolfinu (na rozdíl od večera Plácida Dominga ve Sportovní hale) tuto potřebu skutečně naplnilo.*

[... 10 vět ...]

b. Závěr {bridging, typ=WHOLE_PART na „vystoupení" v *a.*} *patřil španělským zarzuelám.*

III.5.1.1.3. Hraniční případy u asociační anafory typu PART

III.5.1.1.3.1. Hraniční pásmo s nezaznamenáním žádného vztahu

Na pomezí mezi označováním asociační anafory typu PART a nezaznamenáním žádného vztahu stojí páry typu *město – škola* nebo *léčebna – postel* v (27)–(28). Předpokládané části jsou poměrně hluboko „vloženy" do použitého v textu nadřazeného pojmu. Použijeme-li otázkový test z III.5.1.1., také dospějeme k rozhodnutí tyto vztahy nezaznamenávat. Jsou však důležité pro koherenci textu a mezi výrazy je zřejmá souvislost meronymického typu. Nicméně jsme se rozhodli tyto vztahy neanotovat, protože pravidla jejich vymezení jsou příliš vágní a vedou k nutnosti zaznamenání další otevřené množiny vztahů mezi blízkými podtypy, které nejsme schopni z časových a formálně-klasifikačních důvodů uchopit.

(27) *V protidrogové léčebně v Seattlu nezůstala poprvé za dlouhou dobu ani jedna postel* {žádný koreferenční vztah} *volná.*

(28) *V obci Košťany na Teplicku ještě chlapci ani nebyli, ale místní již dali dohromady petici: „My rodiče dětí základní školy Košťany* {žádný koreferenční vztah} *protestujeme proti umístění ubytovny pro potrestané německé chlapce."*

III.5.1.1.3.2. Hranice s asociační anaforou typu FUNCT

V případě vztahu mezi místem a objektem, který může být v daném místě jenom jediný, se vztah mezi výrazy dostává na pomezí mezi asociační anaforu typu PART a FUNCT. Srov. vztah mezi *Mnichov* a *zámek* v (29):

(29) *V Mnichově jsou muzea* {žádný koreferenční vztah} *a galerie* {žádný koreferenční vztah} *se vzácnými obrazy* {žádný koreferenční vztah}, *částečně jsem je navštívil a zhlédl překrásný královský zámek Nymphenburg* {žádný koreferenční vztah}.

III.5.1.1.3.3. Hraniční pásmo s asociační anaforou typu SUBSET

Hraniční pásmo s asociační anaforou typu SUBSET je velice široké a je vyřešeno ve prospěch SUBSET (podrobněji viz III.5.1.2.).

III.5.1.2. Vztah SUBSET mezi množinou a podmnožinou/prvkem množiny (SUB_SET a SET_SUB)

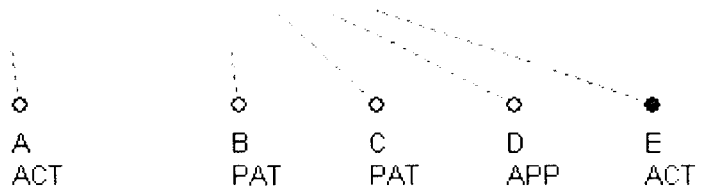
Vztah SUBSET mezi množinou a podmnožinou/prvkem množiny je oboustranný. Sledujeme-li lineární pořadí výrazů v textu, máme následující možnosti:

- SUB_SET – levý člen vztahu referuje na podmnožinu (prvek) množiny referentu pravého členu páru;
- SET_SUB – levý člen vztahu referuje na množinu, pravý člen páru referuje na podmnožinu (prvek) této množiny.

Vzorové příklady vztahu SUBSET jsou např. *Mušketýři – Athos – Porthos – Aramis*; *nápoje – pivo – limonáda – minerálka – cola; motýli – červení – bílí; semináře – první seminář – poslední seminář* apod.

Graficky vztahy typu SUB_SET a SET_SUB lze představit následujícím způsobem (obr. č. 26):

A. Vztah SUB_SET



B. Vztah SET_SUB

Obrázek č. 26: Vztahy typu SUB_SET a SET_SUB

Srov. také následující příklady (30)–(33) z PDT:

- vztah SUB_SET. V prototypických příkladech anaforický člen odkazuje na více antecedentů v předchozím kontextu.

(30) *a. Nejvíce se na tom podílel resort Ministerstva financí ČR – a to formou daňových úlev ve výši zhruba 7.5 miliardy korun.*
b. Další podpora byla poskytnuta Ministerstvem zemědělství – přibližně 2.8 miliardy korun, Ministerstvem práce a sociálních věcí – kolem 178 milionů korun, a Ministerstvem hospodářství – asi 1.2 miliardy korun.
c. V rámci rozpočtové podpory poskytují ministerstva {bridging, typ=SUB_SET na „Ministerstvo financí ČR" v *a.*, „Ministerstvo zemědělství", „Ministerstvo práce a sociálních věcí" a „Ministerstvo hospodářství" v *b.*} *malým a středním podnikatelům zvýhodněné informační služby a poradenskou činnost buď přímo, nebo prostřednictvím specializovaných institucí.*

Srov. také vztah SUB_SET v (31)a–b, kde antecedenty vztahu jsou adjektiva:

(31) *a. Pavel Vondráček: Termín převýchova znám pouze z nacistického [slovníku] a komunistického slovníku.*
b. Na převýchovu se pokud vím, posílali ti, kteří měli podle těchto zrůdných režimů

{bridging, typ=SUB_SET na „nacistický" v *a.*, bridging, typ=SUB_SET na „komunistický" v *a.*} *nevhodný původ.*

- vztah SET_SUB.

Ačkoliv v (32)b anaforický podstrom obsahuje výraz *část*, nejde o typ část – celek, ale o podmnožinu množiny poslanců:

(32) *a. Pokud tedy zrovna nesedí na svém minikřesle v jednací síni, jsou poslanci nuceni pobývat buď ve svých klubech, nebo postávat či posedávat po chodbách.*
b. Nelze se pak ani divit, že část zákonodárců {bridging, typ=SET_SUB, na „poslanec"} *zvolí příjemnější variantu a odchází úřadovat do suterénní restaurace zvané dolní sněmovna.*

Srov. také prvky množiny {masné speciality} v (33)a–b:

(33) *a. Nyní mám firmu, která vyrábí více jak dvě tuny masných specialit denně, mám pět obchodů na severu Moravy.*
b. Firma produkuje na padesát sortimentních druhů párků {bridging, typ=SET_SUB, na „specialita" v *a.*}, *klobásek* {bridging, typ=SET_SUB, na „specialita" v *a.*}, *salámů* {bridging, typ=SET_SUB, na „specialita" v *a.*}, *vyjma trvanlivých.*

III.5.1.2.1. Vztah SUBSET uvnitř jedné věty

Vztah SUBSET se relativně často vyskytuje mezi výrazy uvnitř jedné věty. Označujeme-li důsledně všechny tyto vztahy, některé anotované věty budou v grafickém znázornění přeplněny anaforickými šipkami, což vypadá poněkud pleonasticky. Na druhé straně, pokud vztah množina – podmnožina není zachycen v syntaktické struktuře věty, jeho anotace uvnitř jedné věty není sémanticky méně cenná než v různých větách a měly bychom je anotovat.

Můžeme formulovat následující **pravidlo o postupu při anotaci asociační anafory typu *SUBSET uvnitř jedné věty:***[112]

[112] Toto pravidlo anotace bridging vztahů typu SUBSET uvnitř jedné věty se však (pravděpodobně) nedodržuje v praktické anotaci zcela důsledně.

U vět, ve kterých jsou mezi výrazy vztahy typu SUBSET, postupujeme následujícím způsobem:

1. Pokud výrazy referující na podmnožiny (prvky množiny) jsou členy jedné koordinační konstrukce, ověříme možnost identické textové koreference na spojovací výraz (na spojku, interpunkční známenko) – viz III.4.2.4.1.
2. Pokud uzel s významem podmnožiny (prvku množiny) je přímý potomek uzlu množiny a má funktor MAT, PAT, APP nebo RSTR, vztah asociační anafory mezi nimi neoznačujeme – viz III.5.2.3.
3. V jiných případech asociační anaforu mezi výrazy označíme typem SUB_SET nebo SET_SUB podle lineárního pořadí výrazů v textu. Pokud jde o koordinační konstrukci, šipka asociační anafory povede od konjunktu (resp. na něj).

Srov. následující příklady (34)–(36):

(34) *Jsou mezi nimi například studenti vysokých škol, herečka, kunsthistorik, učitelka,* {bridging, typ=SET_SUB od čárky na `#PersPron`} *psycholožka* .

(35) *Na rozdíl od dobře vybaveného FS dnes nikdo z téměř dvou stovek poslanců kromě předsedy a místopředsedů sněmovny a* {bridging, typ=SET_SUB od spojky na „poslanec"} *šéfů jejích výborů nemá svou kancelář, pracovní stůl, židli a telefon.*

(36) *Proti dřívějšku se však zase objevili noví zájemci o umění z řad podnikatelů, bank, spořitelen a* {bridging, typ=SET_SUB od spojky na „zájemce"} *realitních kanceláří.*

III.5.1.2.2. Asociační anafora typu SUBSET u generických NP a deverbativ

V případě nespecifické (především generické) reference uvažování o vztazích typu SUBSET často vyžaduje určitou abstraktizaci. Vztah SUBSET, kde aspoň jedna jmenná fráze má generickou referenci, má poněkud jinou povahu a jiné sémantické vlastnosti, než vztah výrazů označujících množiny konkrétních vybraných v daném diskurzu objektů reality a jejich podmnožin. Přesto považujeme za smysluplné anotovat vztah SUBSET i v případě vztahu mezi generickými NP a deverbativy.

Vyčleňujeme následující podskupiny:

- vztah „obecný případ s generickou referencí – jednotlivý případ". Např. v (37)a–b vztah mezi *nový VW Golf* v (37)a a *nový golf* v (37)b je označen jako asociační anafora typu SET_SUB. Je to logické už proto, že první výskyt bychom anotovali jako textovou koreferenci typu NR s příp. NP *Nové golfy*, se kterým by vztah SUBSET v (37)b byl nesporný.

(37) *a. Nový VW Golf je vybaven motorem o síle...*
b. Dostali jsme možnost se novým golfem {bridging, typ=SET_SUB na „Golf" v *a.*} *projet.*

- vztah mezi generickými NP, vyjadřujícími podkategorie pojmů. Srov. *žvýkačka* a *bublinová žvýkačka* v (38)a–b:

(38) *a. I když konzervativní Anglie jeho čin odsoudila , guma se zde chytila a Británie se pro žvýkačku stala bránou do Evropy.*
b. Ještě jeden milník si zaslouží zmínku – zrod bublinové žvýkačky {bridging, typ=SET_SUB, na „žvýkačka" v a.}.

Srov. také (39)a–c – (40)a–c:

(39) *a. Firmě dlouho trvalo, než prosadila u německého producenta dělení vodoměrů.*
b. Němci to nechápali.
c. Sami porouchané vodoměry {bridging, typ=SET_SUB, na „vodoměr" v *a.*} *vyhazují celé, protože to je pro ně malá položka.*

(40) *a. Hodnota obrazu však nezávisí jen na autorovi.*
b. Oleje {bridging, typ = SET_SUB, na „obraz" v *a.*} *jednoznačně v čele.*
c. Nejdražší jsou olejomalby {coref_text, typ = NR, na „olej" v *b.*}, *následují tempery a* {bridging, typ = SET_SUB od spojky „a" na „obraz" v *a.*} *pastely, až o polovinu levnější než olejomalba* {coref_text, typ = NR, na „olejomalba"} *může být kresba či* {bridging, typ = SET_SUB od spojky „či" na „obraz" v *a.*} *grafika téhož autora.*

- vztah mezi generickými NP, kde jeden z členů vztahu označuje poměr, počet, procenta, část peněz ve vztahu k celé částce apod. Srov. (41)a–b – (42).

(41) *a. Českobudějovického výrobce školních a kancelářských potřeb opravňuje k optimistickým prognózám velký růst zahraničních zakázek.*
b. Firma dodává na export víc než 60 procent zboží {bridging, typ=SET_SUB, na „potřeba" v *a.*}.

(42) *a. Při výběru pojišťovny jsme zvažovali, kolik by musela zaplatit ročně na pojistném, zda by se mohla připojistit na úraz, zda by byla okamžitě po uzavření pojistné smlouvy pojištěna na sjednanou pojistnou částku a konečně zda si bude moci v případě náhlé potřeby vypůjčit větší sumu peněz z dosud zaplaceného pojistného* {bridging, typ=SET_SUB, na „pojistné"}, *aniž by to mělo vliv na výši pojistné částky.*

- vztah mezi vloženými deverbativy a abstraktními NP. Srov. (43)–(44):

(43) *a. Nová striktní omezení vlády SR proti českým exportérům*
[... 7 vět ...]
b. Jistotu v tomto směru dávají nejnovější kroky vlády SR, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku {bridging, typ=SET_SUB na „omezení" v *a.*} *na zboží zahraniční provenience.*

(44) *a. Jejich vysílače dosud pokrývají signálem programu ČT 2 méně než polovinu území republiky.*
b. Na moravsko-slovenském pomezí je řada míst, kde nezachytí ani první program České televize {bridging, typ=SUB_SET na „ČT 2" v *a.*}.

- vztah „množina konkrétních vybraných objektů – jednotlivec s nespecifickou referencí":

(45) *a. [volontéři] Absolvovali školení v první pomoci pro člověka v nouzi . [...]*
b. Když dítě zavolá, dostane buď radu hned, nebo si s ním volontér {bridging, typ=SET_SUB na „volontér" v *a.*} *domluví další hovor.*

III.5.1.2.3. Hraniční případy u asociační anafory typu SUBSET

III.5.1.2.3.1. Hranice mezi asociační anaforou typu SUBSET a PART

Ve velké části anotovaných vztahů nejde jednoznačně rozhodnout mezi anotací asociační anafory typu SUBSET, nebo PART, resp. jediná skutečnost, o kterou se může opřít anotátor při rozhodnutí mezi těmi dvěma typy, je počitatelnost odpovídajících substantiv. Nejčastěji ambiguita nastává u párů výrazů s generickou referencí, u deverbativ a jmen s abstraktním významem, kde rozdíl mezi částí celku a podmnožinou množiny je často neutralizován. Srov. následující příklady (46)–(47) a všechny příklady z předchozí kapitoly (37)–(45):

(46) *Jeho hlavní výhodou by mělo být lepší napojení na televizní přenosovou techniku: zatímco dnes přenosové vozy* {bridging SUBSET nebo PART na „technika"} *blokují parkovací prostor před starou sněmovnou , v budoucnu zajedou do Thunovské a kabely* {bridging SUBSET nebo PART na „technika"} *se snadno spojí s tiskovým centrem.*

(47) *Ročně by tedy zaplatila na pojistném, včetně úrazového připojištění* {bridging SUBSET nebo PART na „pojistné"}, *4104 korun.*

V obecném případě při rozhodování mezi SUBSET a PART platí následující **pravidlo o preferenci asociační anafory typu SUBSET**:

Není-li rozhodování mezi asociační anaforou typu SUBSET a PART jednoznačné, ambiguitní vztah anotujeme vždy jako SUBSET.

Do asociační anafory typu SUBSET se tedy dostávají všechny ambiguitní případy generických NP a deverbativ (viz III.5.1.2.2.)

Poněkud jinak vypadá situace s příklady (48)–(50). V (48) výraz *zahraničí* v (48)a můžeme zaměnit na synonymickou NP *zahraniční státy* nebo *země,* a v tom případě *Západní Německo* v (48)b se přirozeně interpretuje jako podmnožina zahraničních států. Na druhé straně však vztah mezi územím a částí území anotujeme běžně jako PART a také v daném příkladě *zahraničí* se chápe spíše jako nečlenitelný celek než jako množina. Přikláníme se k názoru, že v daném případě si můžeme dovolit anotaci asociační anafory typu PART.

(48) *a. Cestování se značně uvolnilo až do podzimu 1969, kdy začal být omezen výjezd našich občanů do zahraničí.*

b. Vzpomínám na takzvané zelené hranice zcela bezbariérové a na dosud nevídanou blahovůli zahraničních a našich celních a policejních orgánů už na jaře 1968, tedy ještě chvíli před záborem, kdy jsme jeli autem na výlet do Západního Německa {bridging, typ = WHOLE_PART, na „zahraničí"} *s Hankou Bělohradskou a Dušanem Hamšíkem.*

V (49)–(50) antecedentní výrazy *text* a *knížka* se, podobně jako v (48), teoreticky dají představit jako množina: text je z úzce formálního hlediska množina vět, knížka – množina kapitol. Na druhé straně není vyloučena a je intuitivně přijatelnější interpretace, podle níž je věta neodlučitelnou částí textu a kapitola – knížky. V podobných případech bychom se spíše přikláněli k druhé, intuitivnější variantě.

(49) *a. Když ho smrt překvapila u psacího stolu, revidoval právě text Prezidentské adresy, kterou pronesl několik dní před tím v Americké ekonomické asociaci.*

b. Poslední věta {bridging, typ = WHOLE_PART na „text" v *a.*} *kterou v životě napsal, zněla: Stagnacionisté se mýlí v diagnóze důvodu, proč by kapitalistický proces měl stagnovat.*

(50) *a. Tak je i knížka koncipována.*

b. V každé kapitole {bridging typ = WHOLE_PART na „knížka" v *a.*} *se mluví o určitém problému, uvádíme jak je rozsáhlý, kolik dětí je jím postiženo a co dělat.*

Nicméně, (48)–(50) představují pouze jednotlivé příklady. Při anotaci ve velkém PDT se orientujeme na výše uvedené pravidlo o preferenci typu SUBSET.

III.5.1.3. Vztah FUNCT mezi entitou a unikátní funkcí na této entitě (P_FUNCT a FUNCT_P)

Vztah FUNCT se označuje mezi dvěma uzly (podstromy) v případě když jedna entita vykonává unikátní funkci v rámci jiné entity. Typické příklady vztahu jsou páry *trenér – mužstvo, premiér – vláda, firma – ředitel, akce – organizátor* apod. Entita ve funkci nemusí být pouze jedinec, ale také unikátní úřad, typ *vláda – stát, parlament – stát, národní banka – stát, magistrát – město* apod.

Podobně jako u vztahů SUBSET a PART, vztah FUNCT je oboustranný. Sledujeme-li lineární pořadí výrazů v textu, máme následující možnosti:

- FUNCT_P – levý člen vztahu je unikátní funkce, pravý člen páru je objekt, na kterém je tato funkce definována (typ *trenér – mužstvo*);
- P_FUNCT – levý člen vztahu je objekt, na kterém je definována funkce označena pravým členem páru (typ *mužstvo – trenér*)

Srov. příklady:

- FUNCT_P *ministr financí – ministerstvo financí*:

(51) *Členové parlamentem jmenovaného devítičlenného prezidia FNM jsou: ministr privatizace J. Skalický (předseda), jeho náměstek Jaroslav Jurečka, náměstek ministra hospodářství Václav Kupka, náměstek ministra obchodu a průmyslu Radomír Sabela, viceguvernér České národní banky Pavel Kysilka, náměstek ministra financí Miroslav Téra, předseda FNM T. Ježek, generální ředitel státního podniku Rapid Čestmír Čejka a Pavel Štěpánek z ministerstva financí* {bridging, typ=FUNCT_P, na „ministr financí"}.

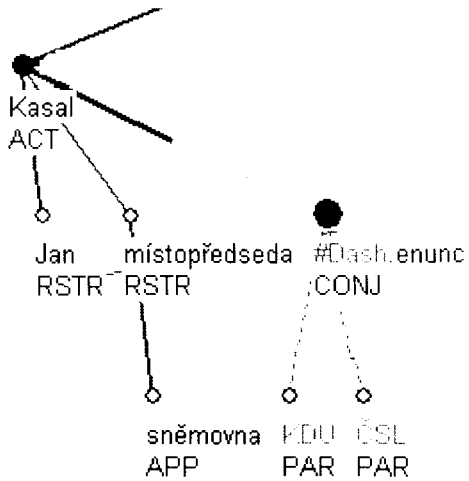
- P_FUNCT v párech *stát – vláda* v (52)a–b a *společnost – generální ředitel* v (53)a–b:

(52) *a. Na přímou podporu podnikání vydá letos stát přibližně 1,8 procenta hrubého domácího produktu.*
b. Tuto skutečnost jednoznačně konstatuje ministr hospodářství Karel Dyba v analýze, kterou předložil vládě {bridging, typ=P_FUNCT, na „stát"}.

(53) *a. Společnost zaměstnává přes dva tisíce zaměstnanců, to je po propuštění důchodců a brigádníků prakticky stejně jako před pěti lety.*
b. S generálním ředitelem Miloslavem Handlem {bridging, typ=P_FUNCT, na „společnost" v *a.*} *jsme hovořili o tom, co se za těmito výsledky skrývá.*

Vztah FUNCT neoznačujeme v případě, že členy vztahu jsou sestry nebo přímí potomci, přičemž závislý uzel má funktor APP. Srov. žádný vztah FUNCT v (54)–(55):

(54) *Místopředseda* {žádný koreferenční vztah} *sněmovny* {žádný koreferenční vztah} *Jan Kasal (KDU–ČSL) uvedl pro LN, že s nesnázemi přijímá skutečnost, když někdo během rozehrané partie mění pravidla hry.*



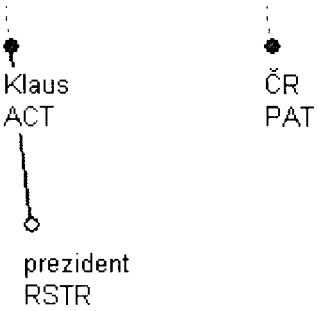
Obrázek č. 27: Vztah FUNCT

(55) *Pokusili jsme se telefonicky kontaktovat prezidenta* {žádný koreferenční vztah} *AC Sparta Praha – fotbal, a. s., Petra Macha, ale jeho vyjádření, zda vyhoví asociaci, jsme nezískali.*

III.5.1.3.1. Označovaní vztahu FUNCT v párech typu *prezident Klaus – ČR*

V konstrukcích, kde se označování funkce skládá z názvu vykonávané funkce a osoby ji vykonávající (*prezident Klaus, ředitelka paní Hlaváčová* apod.), je předmětem označování vztahu asociační anafory název funkce, nikoliv jméno osoby tuto funkci vykonávající, čili v *prezident Klaus – ČR* označíme vztah FUNCT_P mezi *prezident* a *ČR*, nikoliv mezi *Klaus* a *ČR* (srov. obrázek č. 28).

Toto řešení je poněkud komplikované z hlediska syntaktické struktury tektogramatického stromu a navazování vztahů s jinými koreferenčními řetězci. Výraz označující funkci je většinou uzel s funktorem RSTR, závislý na uzlu označujícím právě jméno osoby tuto funkci vykonávající, který označovat nechceme. Přesto však jsme se pro takové řešení rozhodli, a to z následujících důvodů:



Obrázek č. 28: Vztah FUNCT

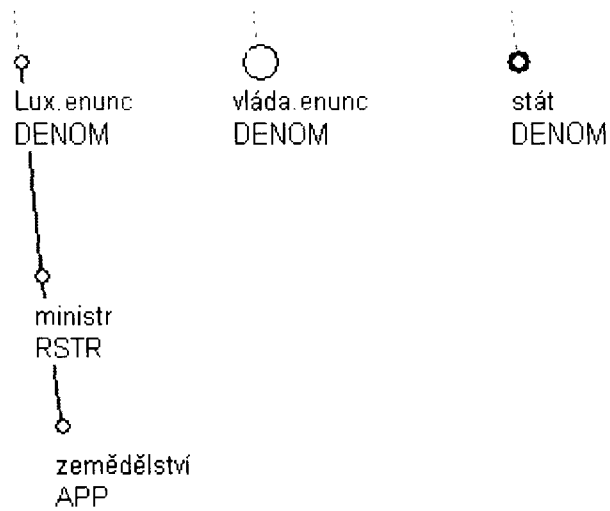
1. Rozhodnutí anotovat FUNCT na jména osob nutí anotátora použít výrazně více znalostí mimojazykové reality, než je třeba pro anotaci souvislého textu.
2. Rozhodnutí anotovat FUNCT na jména osob sníží mezianotátorskou shodu.
3. U vztahu asociační anafory nemusíme dodržovat řetězec, takže není třeba vést šipku násilně na řídící uzel. Označení funkce není v textu vzdáleno od jména osoby tuto funkci vykonávající, a, v případě nutnosti, může být jednoduše automaticky vyhledáno.
4. Asociační anaforu se snažíme pokud možno označovat u výrazů, které mají význam vykonavané funkce ve své lexikální sémantice (*prezident, ministr* apod.) – to pomůže při případném budoucím automatickém zpracování asociační anafory.

Pokud v páru [jméno osoby – objekt, na kterém ta osoba vykonává unikátní funkci] není název vykonávané funkce uveden a odvodíme ho pouze na základě znalosti světa, asociační anaforu typu FUNCT neoznačujeme.

III.5.1.3.2. K otázce hloubky „vloženosti" funkce ve vztazích typu *ministr zemědělství – vláda – stát*

Podobně jako u vztahu PART, v některých případech asociační anafory typu FUNCT je aktuální otázka hloubky „vloženosti" funkce na té či oné entitě. Tak např. vztah FUNCT v párech *ministr zemědělství – vláda* a *vláda – stát* je samozřejmý. Pokud však máme kontext, kde se používají pouze NP *ministr zemědělství* a *stát,* je otázka, zda *ministr* může být ve vztahu FUNCT i k celému státu. Protože podobné situace se opakují v textech PDT relativně často, domluvili jsme se na následujícím postupu:

1. Jsou-li přítomné všechny tři složky vztahu typu *ministr zemědělství – vláda* a *vláda – stát,* jde o ideální situaci a anotujeme dva vztahy FUNCT postupně za sebou, jak je zobrazeno na obrázku č. 29.

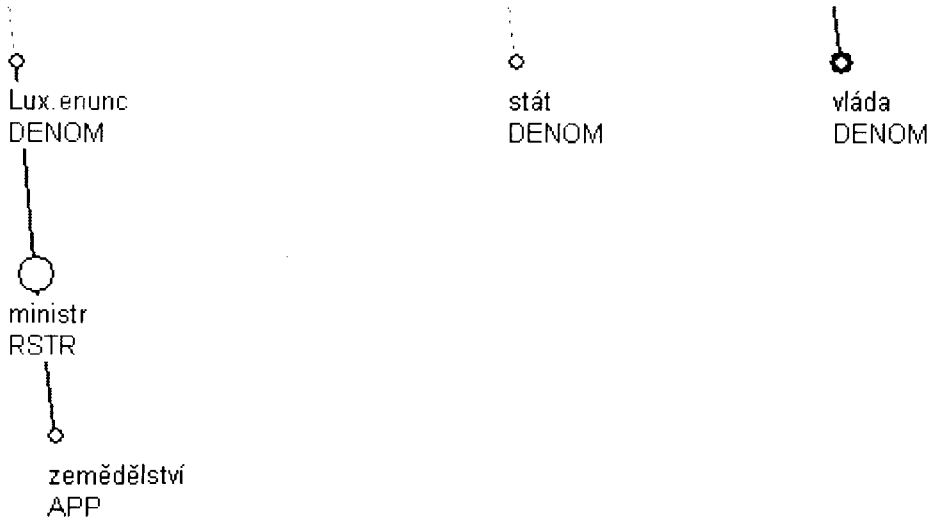


Obrázek č. 29: Vztah FUNCT: hloubka „vloženosti"

2. Pokud chybí prostřední člen, zaznamenáváme pouze vztah FUNCT, jak je zobrazeno na obrázku č. 30.

Obrázek č. 30: Vztah FUNCT: hloubka „vloženosti"

3. V případě, že kontext z 2. pokračuje, a vyskytne se v něm chybějící prostřední člen, stanovili jsme poněkud přebytečnou anotaci: abychom se nemuseli vracet zpět a mazat šipku asociační anafory vedenou jako v 2., povedeme od posledního uzlu dva vztahy typu FUNCT – P_FUNCT na *stát* a FUNCT_P na *ministr zemědělství*, jak je zobrazeno na obrázku č. 31.



Obrázek č. 31: Vztah FUNCT: hloubka „vloženosti"

III.5.1.3.3. Hraniční případy u asociační anafory typu FUNCT

III.5.1.3.3.1. Hranice asociační anafory typu FUNCT a SET

Rozdíl FUNCT a SET se zakládá na unikátnosti funkce vykonávané referentem označeným jako FUNCT nad referentem označeným jako P. Proto např. nespecifikovaný vztah *ministr – vláda* anotujeme jako SET, zatímco *premiér – vláda* – jako FUNCT.

(56) *S oznámeným snížením horní sazby daně z přidané hodnoty a daně z příjmu právnických osob o jedno procento vláda v zásadě souhlasí, a jak po jednání uvedl premiér* {bridging, typ=P_FUNCT, na „vláda"} *Václav Klaus, nelze v tomto bodě očekávat významnější změny.*

Avšak u některých entit unikátnost funkce není relevantní. Srov. v (57)a–b vztah mezi *policie* a *policista* je sémanticky blízky FUNCT, i když je zřejmé, že v policii je policistů daleko více než jeden. Na druhé straně ani vztah SUBSET není zcela adekvátní: policie v tomto významu je těžko představitelná jako množina policistů.

(57) *a. J. Skřičil komentoval dále personální situaci v policii.*
b. K neuspokojivému počtu policistů {žádný koreferenční vztah} *na Olomoucku přispívá částečně i nedostatek bytů ve městě a okolí.*

Podobné případy jsou v anotaci zčásti označeny jako FUNCT, zčásti jako SET nebo REST, některé jsou neoznačené. V daných případech nejsme schopni formulovat přesná pravidla

anotace této široké oblasti vztahů a pokládáme za adekvátnější nabídnout možnost neoznačení vztahu.

III.5.1.3.3.2. Hranice asociační anafory typu FUNCT a neoznačením žádného vztahu

V některých případech není snadno rozhodnout, zda-li daný vztah ještě může být zařazen do FUNCT, nebo už je za jeho hranicemi. Srov. (58)–(59), kde vztah FUNCT již neoznačíme:

(58) *Rozhodování podat si žádost o osvojení dítěte není pro manžele vůbec snadnou záležitostí a může znamenat i konec manželství* {žádný koreferenční vztah}, *kdy se jeden z partnerů s tímto zásadním obratem nedokáže vyrovnat.*

(59) *Navíc mnoho nadaných studentů si vybralo po ukončení studií právě mecenáše své školy* {žádný koreferenční vztah} *jako zaměstnavatele.*

III.5.1.4. Vztah CONTRAST sémantického a kontextového protikladu

Vztah sémantického a kontextového protikladu ve velké míře přispívá ke koherenci textu (srov. k tomu tématu např. identifikátory a alternátory u Palka (Palek 1988), klasifikace typů kontextové zapojenosti u Daneše (Daneš 1999) aj.). Kontrastivní vztahy v textu jsou také relativně jednoznačně vyčlenitelné. Jde o určitý sémantický vztah mezi nekoreferenčními entitami, který nebrání tomu, abychom ho mohli zaznamenávat v rámci asociační anafory.

Vztah asociační anafory typu CONTRAST je jednostranný, oba členy páru považujeme za rovnocenné. Je těžko určit prototypické příklady daného vztahu, protože jsou podmíněny především konkrétním kontextem. K uzlům ve vztahu CONTRASTu sice mohou patřit lexikální antonyma, ale nejčastěji to jsou víceslovné entity, jejichž kontrast zaručuje syntaktická pozice ve větě, rozvíjející členy nebo širší kontext.

Srov. příklady (60)–(62):

(60) *a. Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 mil. USD, tj. téměř 2 % HDP.*

b. I když letos a {bridging, typ=CONTRAST od spojky „a" na „vloni" v *a.*} *příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu, prognózujeme, že saldo přesto zůstane kladné ve výši 300 – 600 mil. USD ročně (1 – 1.6 % HDP).*

(61) *a. Téměř každá třetí tuna prodaného hnědého uhlí na tuzemském trhu pochází z Mostecké uhelné společnosti.*
b. Letos to má být 25.5 milionu tun uhlí.
c. Severočeští horníci se musí vyrovnat se stálým poklesem těžby, která má být v roce 2000 až o třetinu nižší, neboť zejména domácnosti, ale i další spotřebitelé přecházejí na ekologické druhy paliv {bridging, typ=CONTRAST na „uhlí" v *b.*}.

(62) *a. Dnes, po rozdělení ČSFR, je jasné, že osud ČR bude stále více spojený s Německem a přes něj s Evropskou unií a osud Slovenska* {bridging, typ=CONTRAST na „osud ČR" v *a.*} *s Ruskem.*

Asociační anaforu typu CONTRAST neoznačujeme v případě, že jmenné fráze jsou potomky uzlu s funktory ADVS (od *adversative*) a CONFR (od *confrontation*), většinou spojky nebo čárky. Tento funktor vztah kontrastu již vyjadřuje. Srov. např. *zisk – zisk* v (63):

(63) *Dočasný podnikatelův zisk bude anulován, ale.*ADVS *trvalý zisk* {žádný koreferenční vztah} *z jeho inovace zůstane zachován společnosti ve formě nižších cen nebo technicky dokonalejších výrobků.*

Srov. také neoznačovaný CONTRAST u nepřímých potomků uzlu s funktorem ADVS v (64):

(64) *Letos by výstavba technického zařízení v sedmi lokalitách stála 120 milionů korun, ale.ADVS můžeme uvolnit jen 80 milionů* {žádný koreferenční vztah}.

Vztah CONTRAST se částečně prolíná s kontrastem (kontextově kontrastivní zapojenost výrazů – hodnota *c* v atributu tfa v TGS)[113] označeným v rámci anotace kontextové zapojenosti. Jak ukazují příklady, některé šipky asociační anafory s poznámkou CONTRAST jsou v TGS označeny jako kontextově kontrastivně zapojené výrazy. Není to však pravidlem. Srov. např. v (65)a–b kontrastivní vztah mezi *minulý rok* a *letos* je zachycen jak v anotaci

[113] Anotace kontextové zapojenosti je provedená ručně na tektogramatické rovině na celém korpusu PDT 2.0. Na základě zapojenosti nebo nezapojenosti výrazu do kontextu se v anotaci rozlišují tři typy výrazů: kontextově nekontrastivně zapojený výraz (hodnota *t* v atributu tfa), kontextově kontrastivně zapojený výraz (hodnota *c* v atributu tfa) a kontextově nezapojený výraz (hodnota *f* v atributu tfa). Podrobněji viz Mikulová a kol. 2005, s. 1054n.

kontextové zapojenosti, tak i v anotaci asociační anafory. Na *letos* se však navazuje další CONTRAST *v roce 1995*, který už v anotaci kontextové zapojenosti zachycen není:

(65) *a. Například minulý rok* {hodnota c v atributu `tfa`} *byla velká poptávka po měřičích tepla, jejichž montáž diktoval zákon.*
b. Letos {hodnota c v atributu `tfa`, bridging CONTRAST na „minulý rok" v *a.*} *se počítá se zájmem o bytové vodoměry, v roce 1995* {hodnota f v atributu tfa, bridging CONTRAST na „rok 1995"} *se očekává poptávka po regulačních prvcích na radiátory.*

V tabulce č. 20 uvádíme několik možných kombinací hodnot atributu `tfa` u elementů spojených vztahem asociační anafory typu CONTRAST. Možné jsou i další kombinace.

příklad s anotovaným vztahem CONTRAST	**hodnoty atributu tfa**
(66) *A přesvědčen jsem ještě o jednom – je třeba mít vysoké cíle a s malými [cíli]* {bridging, typ=CONTRAST, na „cíl"} *se nespokojit.*	**f – f**
(67) *Co se může dospělému zdát zanedbatelnou záležitostí, naroste v dětské* {bridging, typ=CONTRAST, na „dospělý"} *mysli třeba i do tragických rozměrů.*	**c – c**
(68) *Všichni účastníci obchodů na burze by měli podle Jiřího Béra vědět, že bývalý majitel CP má nárok na dividendu v době jeho prodeje, tedy v čase T+2 včetně.* *(15) Pokud dojde k omylu, lze zpětně požádat nového majitele* {bridging, typ=CONTRAST, na „bývalý majitel"}, *aby poukázal peníze správnému majiteli CP.*	**t – t**
(69) *Lidi nežvýkají, to jenom krávy* {bridging, typ=CONTRAST, na „člověk"}.	**c – f**

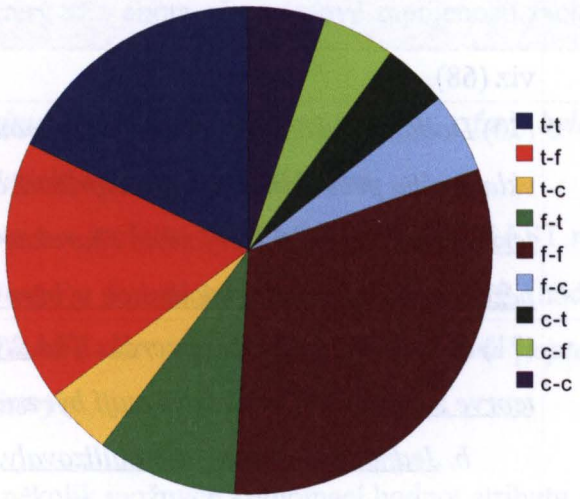
Tabulka č. 20: Vztah CONTRAST a kontextová zapojenost výrazů

Statistický přehled kontextového zapojení výrazů – účastníků vztahu asociační anafory typu CONTRAST uvádíme v tabulce č. 21:

hodnoty atributu `tfa`	**počet**	**příklad**
t – t	111	viz (68)
t – f	108	(70) *Dodal, že ukončení smluv se týká pouze speciálního vkladového produktu, který je nesprávně označován jako terminovaný vklad a který svým charakterem spíše odpovídá běžnému účtu s roční periodicitou výběru zůstatku.*
t – c	28	(71) *a. Nekázeň se podle generála týká útvarů, jež transformace teprve zasáhne, zejména když mají být zrušeny.* *b. Jednotky, které se již stabilizovaly, odvádějí naopak dobré výsledky.*
f – t	58	(72) *a. Že mnohé Němce bude v těchto dnech bolet hlava, s tím se počítalo.* *b. Německé podnikatele však čeká aspoň po několik příštích týdnů bolehlav ze zcela jiných příčin.*
f – f	195	viz (66)
f – c	33	(73) *a. Jen letos zahynulo k 11. červenci ve službě 6 základáků (většinou ve strážní službě) a 1 profesionál.* *b. V době volna či dovolených 19 základáků (zpravidla dopravní nehody) a 8 profesionálů (sebevraždy).*
c – c	32	viz (67)
c – t	27	(74) *Většina západoevropských států se prý obává, že jiná populace v jejich zemích přeroste nad populací původní.*
c – f	30	viz (69)
CELKEM	655	

Tabulka č. 21: Hodnoty atributu tfa a asociační anafora typu CONTRAST

Rozložení hodnot atributu tfa ve vztahu asociační anafory typu CONTRAST ukazuje graf č. 3:



Graf č. 3: Rozložení hodnot atributu `tfa` *ve vztahu asociační anafory typu CONTRAST*

Částečné prolínání asociační anafory typu CONTRAST a anotace kontextového zapojenosti může být pro příp. lingvistický výzkum kontrastivnosti velice zajímavé. Po ukončení anotace rozšířené textové koreference a asociační anafory a po zapracování do PDT diskurzivních vztahů (viz Mladová a kol. 2008, Mladová 2008, Zikánová 2007) bude kontrastivnost zpracována na třech úrovních:

- kontrastivnost na úrovni výrazů (kontextová zapojenost, hodnota *c* v atributu `tfa`);
- kontrastivnost na úrovni podstromů (asociační anafora typu CONTRAST);
- na úrovni propozic (projekt anotace diskurzu).

III.5.1.4.1. Hraniční případy u asociační anafory typu CONTRAST

III.5.1.4.1.1. Hranice mezi asociační anaforou typu CONTRAST a identickou textovou koreferenci

Nejčastější případ problematických CONTRASTů jsou předložkové fráze, jejichž NP-složky jsou koreferenční, ale které se v předložkové konstrukci dostávají do kontrastu (typ *před válkou – po válce*), viz k tomu III.4.2.3.5.2. o předložkových konstrukcích. Na základě principu o preferenci koreference před asociační anaforou (III.5.1.) podobné případy anotujeme jako identickou koreferenci.

Existují však kontexty, kde jmenné fráze, které jsou součástí předložkových frází nejsou v našem smyslu koreferenční (např. *před rokem 1989 – po listopadu 1989*); v takovém případě můžeme daný vztah označit jako asociační anaforu typu CONTRAST. Srov. (75):

(75) *a. Její poslanci se před rokem 1989 scházeli čtyřikrát do roka, odhlasovali vše, co se jim řeklo, a pak se rychle vrátili do svých domovů, kde některým běžel plat koncernových ředitelů a jiným dojiček krav.*
b. Po listopadu 89 {bridging, typ=CONTRAST, na „rok" v *a.*} *se poslancování stalo placenou činností a nároky na jeho vykonávání přiměřeně tomu vzrostly.*

III.5.1.4.1.2. Hranice mezi asociační anaforou typů CONTRAST a ANAF

viz III.5.1.5.1.2.

III.5.1.5. Vztah ANAF anaforického odkazování mezi nekoreferenčními entitami

V přirozených souvislých textech se často setkáváme s příklady explicitního anaforického odkazování (pomocí identifikátorů) na nekoreferenční antecedent. Pro příklady, které nezapadají do již popsaných skupin asociační anafory (typy PART, SUBSET, FUNCT a CONTRAST), jsme zavedli speciální vztah ANAF nekoreferenční anafory.

Vztah asociační anafory typu ANAF je jednostranný: jak plyne z názvu, odkazuje vždy anaforicky dozadu. Prototypické příklady mají bez kontextu slabou exemplifikační sílu (pro určování vztahu je nutno znát referenční charakteristiky výrazů ve výpovědi), jsou to např. *leden – ve stejném období loňského roku, duha – toto slovo, Rakousko přepadlo Maďarsko – v tu dobu* apod.

Asociační anaforu typu ANAF anotujeme v následujících případech:

1. Metajazykové odkazy (odkaz na pojmenování, nikoliv na denotát), diskurzivní deixe v pojetí Fr. Lenze.[114] Srov. např. typické příklady ze SYN2005 (76)–(77):

(76) *Protože tenhle Adolf Hitler nebyl vůdce velkoněmecké říše, ale pták druhu tučňák královský. A to jméno dostal vlastně dodatečně.* (Zábrana, J., Vražda v zastoupení)

[114] Discourse deixis: "reference to linguistic entities rather than to the (identical) referents of antecedent expression" (Lenz 2007, s.71).

(77) *[...] a kolem trůnu duha jako smaragdová. Mari odstrčí židli, na které sedí, a postaví se. „Duha?" Kněz přiloží prst k tomu slovu, aby nezapomněl, kde skončil.* (Ludva, R., Jezdci pod slunečníkem)

Srov. také příklady (78)–(79) z PDT, které nejsou tak jednoznačné, jako uvedené příklady ze SYN2005, ale patří také do asociační anafory typu ANAF. Na anaforičnost druhého členu párů poukazuje výrazná tematičnost a kontextová zapojenost NP *převýchova* v (78)b a *talent* v (79):

(78) *a. Pavel Vondráček: Termín {PAT} převýchova {ID} znám pouze z nacistického a komunistického slovníku.*
b. Na převýchovu {bridging ANAF na „termín"} *se pokud vím, posílali ti, kteří měli podle těchto zrůdných režimů nevhodný původ.*

(79) *a. Co si pod pojmem talent představujete?*
b. Talent {bridging ANAF na „pojem" v *a.*} *je...*

2. Odkaz na čas, ve kterém probíhá antecedentní událost, popsána ne více, než jednou větou (v případě, že antecedentní událost je popsána více větami viz III.7.2.). Anaforické odkazy tohoto typu jsou často implicitně korelační, tj. anaforickou větu můžeme přeformulovat na „právě v té době, kdy se to dělo", nebo „právě v tom okamžiku, kdy to řekl". Tuto skutečnost můžeme použít jako neformální test na identifikaci vztahu ANAF. Srov. (80)–(82):

(80) *a. Tak jako každý Mexičan, i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle, a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku.*
b. Právě v té době {bridging, typ=ANAF na celé *a.*} *přihrála náhoda Santa Annovi do cesty Thomase Adamse, fotografa a především vynálezce všeho druhu.*

(81) *Po chvilce mlčení se Billy otázal: „Je tady už ten druhý?" „Zatím ne." Ale právě v tom okamžiku* {bridging, typ=ANAF na předchozí větu} *se objevil muž, jehož tvář viděl Broderick v dokumentaci Teda Sanderse vedle fotografie Hubbardovy.* (Polák, J., Závody)

(82) *a. Rozbití Varšavské smlouvy bylo jako odseknutí údů od těla.*
b. Od té doby {bridging, typ=ANAF na „rozbití" v *a.*} *se toho mnoho neudělalo.*

3. Odkaz na typově podobný objekt, tzv. „anafora na příznak" (viz Padučevová 2007), často se slovy *takový, podobný, stejný.* Srov. segmenty vět v (83)–(86):

(83) *a. Nic nenasvědčuje tomu, že by parlamentní budova měla sloužit jiným než parlamentním účelům.*
b. Přesto se takové názory {bridging, typ=ANAF na „sloužit" v *a.*} *ozývají.*

(84) *a. Linka bezpečí 855 44 33, kterou od zítřejšího dne provozuje v Praze nadace Naše dítě, byla zřízena především s úmyslem pomoci fyzicky i duševně týraným a pohlavně zneužívaným dětem. (...)*
b. Ptali jsme se několika, jestli by takového kamaráda po telefonu {bridging, typ=ANAF na „linka" v *a.*} *považovali za dobrou věc.*

(85) *a. ...vojáci zde povalili a zdemolovali telefonní budku.*
b. Podobné problémy... {bridging, typ=ANAF na „a" v *a.*}

(86) *a. Ministerstvo se nechystá zakázat žádnou extremistickou organizaci.*
b. ...že by se něco takového {bridging, typ=ANAF na „zakázat" v *a.*} *chystalo.*

III.5.1.5.1. Hraniční případy u asociační anafory typu ANAF

III.5.1.5.1.1. Hranice mezi pronominální textovou koreferencí a asociační anaforou typu ANAF

Často se stává, že se v textu objeví spojení „řekl to X", přičemž zájmeno *to* odkazuje k celé předcházející větě. V rámci předchozí anotace pronominální koreference označujeme takový odkaz jako textovou koreferenci, v některých případech však toto označení neodpovídá skutečnosti a někdy dokonce může způsobovat problémy. Srov. např. (87)a–c:

(87) *a. Kuchyňským nožem ubodal v noci z neděle na pondělí třiačtyřicetiletý J. S. v kuchyni bytu v Pekařské ulici svou o jedenáct let mladší manželku.*

b. LN to sdělil vyšetřovatel Krajského úřadu vyšetřování.

c. Motivem činu, který pachatel sám ohlásil, byly déletrvající manželské neshody.

Výraz *to* v (87)b je anotován jako koreferenční s řídícím slovesem *ubodal* zastupujícím celou větu (87)a. Výraz *čin* v (87)c je pak koreferenční se slovesem *ubodal,* ovšem nikoli s *to* v (87)b. Spojíme-li textovou koreferencí *čin* a *ubodat,* koreferenční vztah automaticky povede k *to,* což neodpovídá skutečnosti. V daném případě jde o odkazování na různé aspekty významu jedné diskurzivní jednotky (v příkladě (87)a–c – celé věty (87)a, na kterou se v (87)b odkazuje jako na sdělenou informaci a v (87)c – jako na uskutečněnou událost), srov. také podobný případ v III.4.2.3.5.2.

Možné řešení pro tuto nesrovnalost je anotovat vztah mezi *to* v (87)b a antecedentní větou v (87)a jako asociační anaforu typu ANAF. Ve skutečnosti to znamená měnit již anotovanou pronominální textovou koreferenci na asociační anaforu typu ANAF. Taková anotace sice vyřeší nekoreferenční řetězce v případech daného typu, má však následující nevýhody:

- jde proti principu o preferenci textové identické koreference před asociační anaforou;
- je to časově náročná změna, vyžadovaná na datech, která jsou již oanotovaná;
- změna coref_text na ANAF není tak jednoznačná, jak se na první pohled zdá. Např. není tak jednoznačné, zda zájmeno *to* v anaforické větě metajazykově odkazuje na výpověď o situaci, nikoli na situaci samu. Např. v uvedeném příkladě (87)b vyšetřovatel Krajského úřadu mohl prostě informovat o situaci, aniž by použil přesně větu (87)a.
- změnou původní anotace bychom výrazně zkomplikovali mezianotátorskou shodu. Kromě konstrukcí s „řekl to" se může vyskytnout také např. „prohlásil to", „oznámil to", „sdělil to", „zmínil se o tom", „naznačil to" apod. Těžko bychom se ve všech případech shodli.

Jako metajazyk bychom mohli podobné odkazy chápat v případě, kdyby předchozí věta byla v uvozovkách jako citace. Avšak i v tomto případě považujeme takové řešení za zbytečnou komplikaci. Proto v těchto případech původní anotaci zájmenné koreferenci na asociační anaforu typu ANAF neměníme.

III.5.1.5.1.2. Hranice mezi asociační anaforou typů ANAF a CONTRAST

Problém výběru mezi vztahy asociační anafory typů ANAF a CONTRAST vzniká především v případě anafory na příznak se slovy *takový, podobný* ale také např. *jiný.* Pokud

anaforická NP je specifikována alternátorem, vybíráme CONTRAST. V ostatních případech anotujeme ANAF. Srov. např. (88):

(88) *...náklady nutného a neodkladného léčení v zahraničí... ...do výše nákladů spojných s takovým léčením na území ČR* {bridging, typ=ANAF na „léčení"}...

III.5.1.5.2. Anaforické odkazování na nevyjádřený antecedent

Určitou lingvistickou zajímavost představují anaforické odkazy na nevyjádřené (ani neelidované) entity. Srov. např. v (89) anaforický odkaz od *daný region* na nevyjádřené neobligatorní místní určení:

(89) *Jde o významný vztah nepřímé úměry – čím vyšší počet živností [v regionu – A.N.], tím relativně nižší nezaměstnanost v daném regionu* {žádný koreferenční vztah}.

V (90)a–b je anaforický odkaz od NP *oba nástupnické státy* na implicitní nevyjádřený antecedent, který se vyvozuje na základě kontextu (rozdělení Československa na Česko a Slovensko). Daný anaforický koreferenční vztah by byl anotován v případě, kdyby syntaktická anotace na tektogramatické rovině obsahovala všechny doplněné aktanty u substantiv na -ní/-tí – koreferenční šipky by vedly na tyto aktanty.

(90) *a. Případ rozdělení Československa budí v zahraničí i po dvou letech pozornost.*
b. Evropa oceňuje nenásilnost rozchodu, považuje oba nástupnické státy {žádný koreferenční vztah} *za součást zóny stability, ale často zapomíná ve výčtu nejbližších kandidátů členství v Evropské unii na jeden z nich, na Slovensko.*

Tyto případy neznačíme, protože pro ně nelze najít antecedent.

III.5.1.6. Vztah REST pro jiné případy asociační anafory

Vztah REST anotujeme v případě, že mezi entitami je zřetelný vztah asociační anafory; speciální kategorie pro tento typ vztahu však není zavedena. Při zaznamenání asociační anafory typu REST dbáme na to, aby tato kategorie nebyla použita jako „sběrný koš", do kterého vřadíme všechny páry výrazů, které mezi sebou nějakým způsobem souvisí. Daná kategorie slouží především jako prozatímní sběrna typologicky podobných párů, které mohou být v

budoucnu vyčleněny do zvláštní skupiny. Na dané etapě do RESTu zařazujeme následující typy vztahů:

III.5.1.6.1. Vztah rodinné příslušnosti

Vztah mezi pojmenováními rodinných příslušníků (*otec – syn, manžel – manželka* apod.). Srov. (91)–(92):

(91) *a. Úzce navazuje na tradici podnikání svého rodu, především dědy.*
b. Od něj {coref_text na „děda"} *získal vnuk* {bridging, typ=REST na „#PersPron"} *výtečné základy, ač sám vystudoval školu zaměřenou na dopravu.*

(92) *Pokud to bude potřeba a dítě k tomu dá souhlas, pozve je* {coref_text na „dítě"} *v doprovodu rodiče* {bridging, typ=REST na „#PersPron"} *nebo jiného dospělého do dětského krizového centra, jež tvoří zázemí linky.*

III.5.1.6.2. Vztah „místo – obyvatel"

Vztah *místo – obyvatel* (*Praha – Pražáci, Rusko – Rusové, Mexiko – Mexičan* ale také u obecných jmen v párech typu *stát – obyvatelé, země – veřejnost* apod.) se v některých klasifikacích vyčleňuje jako zvláštní kategorie. Je přesně vyčlenitelný a zřejmě se podílí na kohezi textu. Srov. např. (93)–(94):

(93) *a. Poté, co byl v roce 1845 jako prezident svržen a na deset let vypovězen na Kubu, vydal se do New Yorku s jedinou myšlenkou – získat zpět vládu nad Mexikem.*
b. Tak jako každý Mexičan {bridging, typ=REST na „Mexiko" v *a.*}, *i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle ....*

(94) *a. Kwasniewski opakovaně zdůraznil, že z cesty zásadních proměn země nelze sejít: pravice a levice se budou přít se středem o tempo změn, ale o základní vzorec reforem není sporu.*
b. Co však je vážné, je nevelký zájem veřejnosti {bridging, typ=REST na „země" v *a.*}, *o věci veřejné.*

Srov. také (95), kde vztah mezi antecedentem a anaforem není úplně zřejmý:

(95) *Keř, kterého si Kolumbus na ostrově Santo Domingo povšiml, je příbuzným řecké mastiky a jeho mízu místní Indiáni* {bridging, typ=REST na „ostrov"} *používali stejně jako Řekové.*

III.5.1.6.3. Vztah typu „autor – dílo"

Vztah mezi autorem a jeho dílem označujeme jako REST v případě, pokud uzel referující na autora není přímým potomkem s funktorem AUTH uzlu s významem díla. Srov. typické:

(96) *Při výběru obrazu bude hrát určitě velkou roli autor* {bridging, typ=REST na „obraz"}.

Žádný vztah však nezaznamenáváme v následující větě (97):

(97) *Krásná, ale nesignovaná krajinka* {žádný koreferenční vztah} *neznámého malíře* {funktor AUTH} *bude určitě hůře prodejná než slabý Slavíček.*

III.5.1.6.4. Vztah „věc – majitel"

Vztah mezi objektem a jeho majitelem označujeme jako REST. Tento vztah však nezaznamenáváme, pokud uzel označující majitele je přímým potomkem s funktorem AUTH uzlu s významem objektu ovládání.

Srov. vztah REST mezi *obraz* a *majitel* v (98):

(98) *Obraz výrazně stoupne na ceně, má-li majitel* {bridging, typ=REST na „obraz"} *doklad o tom , že byl vystaven na výstavě, či je publikován v knize či katalogu.*

III.5.1.6.5. Vztah mezi stejně vyjádřenými nebo synonymními nekoreferenčními NP

Tento vztah je blízký již vyčleněným typům asociační anafory CONTRAST a ANAF, ale některé případy nelze zařadit ani do jedné z těchto skupin. Asociační anaforu typu REST tedy označujeme mezi jmennými frázemi, pokud zároveň splňují následující podmínky:

- NP jsou totožné nebo synonymní;
- NP nejsou koreferenční (třeba jeden člen má generickou, druhý specifickou referenci);
- vztah mezi danými NP se podílí na koherenci textu;

- vztah mezi danými NP nemůže být zařazen do asociační anafory typů CONTRAST nebo ANAF.

Srov. následující příklady:

a) členy vztahu jsou totožná pojmenování:

(99) *a. „Sever Čech má za sebou svízelnou minulost, má před sebou po skončení vlády komunistů novou naději," domnívá se Raimond Strathman, člověk zodpovědný za akci Evangelické diakonie v České republice.*
b. Staráme se o děti, které se ne vlastním přičiněním dostaly do těžké situace.
c. Vždyť i ony mají před sebou novou naději {bridging, typ=REST na „naděje" v *a.*}.

(100) *a. Právě v té době přihrála náhoda Santa Annovi do cesty Thomase Adamse , fotografa a především vynálezce všeho druhu.*
b. Psal se rok 1869 a do hry vstoupila další náhoda {bridging, typ=REST na „náhoda" v *a.*}.

V (101)a–c již není zcela jisté, že se ten vztah má označovat. Hranice mezi označením a neoznačením je v podobných případech poměrně vágní a určuje se pokaždé zvlášť na základě kontextu.

(101) *a. Keř, kterého si Kolumbus na ostrově Santo Domingo povšiml, je příbuzným řecké mastiky a jeho mízu místní Indiáni používali stejně jako Řekové.*
b. Zatímco karibští Indiáni strčili do úst kousek surové gumy v té podobě, jak jej utrhli od kůry, Mayové na poloostrově Yucatán přivedli žvýkání na vyšší úroveň.
c. Mízu {bridging, typ=REST na „míza" v *a.*} *stromu sapodilla sklízeli a upravovali systémem, který se používá dodnes.*

b) členy vztahu jsou synonymická pojmenování:

V příkladě (102)a–b dvě nekoreferenční NP *holky – děvčata* (jde o situaci v různých regionech, tedy holky/děvčata jsou v (102)a a (102)b různé) jsou synonymní, zcela jistě spolu sémanticky souvisí a mají vliv na koherenci textu. Obě NP mají generickou referenci.

(102) *a. Měli časté incidenty s městskou policií: „Díky tomu, že měli dostatek finančních prostředků, byli často opilí, nabalovala se na ně místní mládež, a také si brali do kláštera holky," vzpomíná obecní strážník.*

[... 3 věty ...]

b. I tady si prý chlapci, kteří měli být vychováváni na faře, užívali děvčat {bridging, typ=REST na „holky" v *a.*} *a svobody.*

(103) *a. Stačilo jen razítko na hranicích, celní kontroly jejich orgány nedělaly, náš pas – tedy to, že jsme z Československa, byla sama o sobě průkazná vizitka a vstup na jejich území (a stejně tak i na další) byl hladký.*

b. Naši celníci už nás čekali, viděli nás, jak vyjíždíme od Němců, zastavili jsme, dali štempl {bridging, typ=REST na „razítko"} *a měli jsme jet dál.*

III.5.1.6.6. Vztah „událost – argument"

Vztah mezi substantivně označenou událostí a jejími argumenty je tradičně vymezován v rámcích typů asociační anafory (srov. Clark 1977, Gardent 2003, Chiarchos – Krasavina 2005 aj.). V naší anotaci však nebyl vymezen jako zvláštní kategorie z důvodu své relativně nízké frekvence v textech PDT, ale také proto, aby nás implicitně nevedl k označování asociační anafory mezi slovesem a jeho aktanty, neboť tento vztah je již zaznamenán v tektogramatické struktuře věty. Pokud se takový typ vyskytne v anotovaných datech, označíme ho jako asociační anaforu typu REST. Jde o vztahy typu *podnikání – podnikatel, podnik*; *spor – účastník konfliktu, zpěv – zpěvák, píseň* apod. Srov. např. *podnikání – podnikatel* v (104)a–b a *prostitutky – prostituce* v (105)a–b:

(104) *a. Relativně tak stát vynakládá na tržně konformní podporu malého a středního podnikání přibližně 1.6 – 1.8 % hrubého domácího produktu.*

b. V rámci rozpočtové podpory poskytují ministerstva malým a středním podnikatelům {bridging, typ=REST na „podnikání" v *a.*} *zvýhodněné informační služby a poradenskou činnost buď přímo nebo prostřednictvím specializovaných institucí.*

(105) *a. V ČR bývají prostitutky zadržovány zpravidla jenom kvůli ověření totožnosti.*

b. Zákon o prostituci {bridging, typ=REST na „prostitutky"} *se u nás teprve*

připravuje.

Vztah mezi slovesem a jeho argumenty neoznačujeme, čili např. vztah *podnikat – podnikatel* již nebude zaznamenán.

III.5.1.6.7. Vztah „objekt – velmi typický instrument"

Srov. (106)a–d:

(106) *a. Začal jsem, řekněme, jako provazochodec.*
b. Lidé chodili po zemi, já nějakých dvacet centimetrů nad ní.
c. Klidně jsem mohl seskočit a dál dělat ve státním podniku, nic by se nestalo.
d. Ale začal jsem lano {bridging, typ=REST na „provazochodec" v *a.*} *zvedat a seskočit už nebylo možné.*

III.5.1.6.8. Jiné možné vztahy, o kterých jsme uvažovali

Pro anotaci některých případů by se jako relevantní jevila kategorie asociační anafory označující vztah společného členství v množině. Jde o situaci kohyponymie, kdy se do textu uvádí entita, která je kontextově (nikoli syntakticky) souřadná s jinou entitou, např. *trh zlata – trh ropy, poslanec A – poslanec B* apod. Pokud mezi těmi entitami není vztah kontrastu ani se nevyskytuje zahrnující pojem, se kterým bychom mohli obě entity propojit vztahem SUBSET nebo PART, nemáme pro zaznamenání jejich vztahu žádnou vymezenou kategorii. Avšak zavedení takové kategorie by znamenalo výraznou komplikaci pro anotační práci: měli bychom stanovit preference vztahu SUBSET a PART a v případě, že kontext s již označenými tímto novým vztahem entitami bude pokračovat nadřazeným pojmem, budeme muset tyto vztahy smazat a nastavit SUBSET. Kromě toho, tento významový vztah se nezdá ani těsný, resp. je volnější než některé jiné vztahy, které se neanotují. Výrazné případy toho vztahů, pokud neobsahují kontrast, mohou být označeny jako asociační anafora typu REST.

III.5.2. K omezení počtu vztahů asociační anafory

Jak jsme již naznačili v III.4.2.3. podstata koherence textu je jev velice složitý a mnohostranný. Kategorie vztahů asociační anafory zachycuje pouze malou část skutečných vztahů v textu, a to také pouze formálně, tedy pro strukturu textu ne úplně reprezentativně.

Jiná možnost zachycení koherenčních vztahů v textu je anotace hypertematických řetězců. Podobné projekty na velkém korpusovém materiálu také existují (srov. landmark coding, MapTask) a jsou dokonce zpracovatelné automaticky.

Pokus o označení hypertematických vztahů byl proveden na jednom souboru PDT (cmpr9410_028.t.). Je to text o 139 větách obsahující dvě průběžná hypertémata – umělecká a peněžní. V (107) je uveden jeden hypertematický řetězec z prvních 60 vět souboru cmpr9410_028.t.gz:

(107) *peníze – investování – cena – koupit – prodělat – prodat – pořídit – stát – zhodnocení – poptávka – trh – nabídka – kupce – rozpočet – hodnota – nejdražší – nejlevnější – 9 tisíc – zaplatíte – kupující – ceny stouply – trojnásobek – vyplatí se – prodej atd.*

Na jedné stráně je škoda v naší anotaci takové výrazy nepropojovat. Identickou koreferencí je však všechny spojit nemůžeme, protože referují pokaždé k jiným objektům. Co se týče asociační anafory, těch málo typů, které jsme vymezili, a anotační pravidla ohledně anotovaných slovních druhů nestačí pro propojení všech elementů hypertematického řetězce: některé typy budeme muset vynechat, pro některé nenalezneme vztah atd. Také mezianotátorská shoda bude v takovém případě minimální.

Variantním řešením je zavést zvláštní vztah – hypertematický – který by takové řetězce propojoval. Avšak ruční provedení takové práce je velice časově náročné. Do hypertematického řetězce mohou patřit všechny autosémantické lexikální prvky, bez zřetele k jejich slovnímu druhu a referenční platnosti. Někdy, a to poměrně často, se v jedné větě vyskytuje několik uzlů daného sémantického pole, které budou mezi sebou propojeny. Tyto vztahy se pak zase složitým způsobem proplétají do existujících vztahů textové koreference a asociační anafory. Ve výsledku dostaneme text do takové míry přeplněný šipkami, že už těžko může být použitelný k nějakému konkrétnímu účelu, protože jsou v něm v podstatě téměř všechny uzly mezi sebou propojeny. Kromě toho v dané fázi vývoje počítačové lingvistiky již existují automatické statistické způsoby vyhledávaní hypertémat článků.

Z těchto důvodů hypertematické vztahy ve stávající anotaci neoznačujeme.

Abychom předešli chaotičnosti a subjektivnosti anotace asociační anafory, vypracovali jsme několik pravidel, které omezují počet vztahů asociační anafory v naší anotaci.

III.5.2.1. Preference koreference

Pokud můžeme vybírat, zda odkázat na (třeba i vzdálenější) uzel v textu identickou koreferencí, nebo vztahem asociační anafory, vždy volíme identickou koreferenci (viz princip preference koreference v III.1.5.).

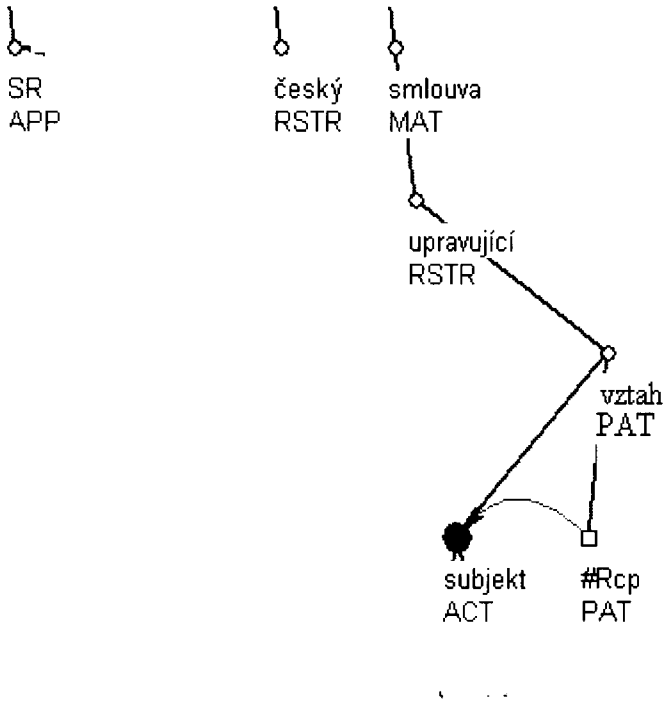
III.5.2.2. Ne více než jeden vztah od jednoho uzlu

Pokud od daného uzlu již vede jeden typ vztahu asociační anafory, nehledáme další. Srov. např. v (108)a–c vidíme dva typy vztahů – dva vztahy SUBSET (*Podání intelektuálské – tvář* a *Podání pragmaticky obchodní a ekonomické – tvář*) a CONTRAST v páru *Podání intelektuálské* a *Podání pragmaticky obchodní a ekonomické*. Jakmile jsme stanovili, že pro asociační anaforu vybíráme vždy jenom jeden vztah, zaznamenáváme pouze asociační anaforu typu SUBSET, protože v daném příkladě a) nese více informace, b) v kontextu se s ním setkáme dříve, než s CONTRASTem:

(108) *a. Tento postoj má více tváří.*
b. Podání intelektuálské {bridging, typ=SET_SUB, na "tvář" v *a.*} *pochází z pochybování o veškeré realitě včetně sebe samého a ústí v postmodernistický relativismus a neschopnost zaujmout pevný jednoznačný postoj.*
c. Podání pragmaticky obchodní a ekonomické {bridging, typ=SET_SUB, na "tvář" v *a.*} *jednostranně preferuje krátkodobé, praktické potřeby, tedy potřeby, zaměřené na současnost a bezprostřední budoucnost.*

Počet vztahů jednoho typu od jednoho uzlu neomezujeme. Tento jev je typický, především u asociační anafory typů SUBSET a PART (protože podmnožin u množiny a částí u celku může být mnoho), je možný také u FUNCT (viz případ *ministr – vláda – stát* v III.5.1.3.). U vztahů REST a CONTRAST se z pochopitelných důvodů nevyskytuje. Srov. dva vztahy asociační anafory typu SUBSET, od *subjekty celního soustátí* v (109)b k *SR* a *český* v (109)a a jejich vztah zobrazený na obrázku č. 32:

(109) *a. Nová striktní omezení vlády SR proti českým exportérům*
b. Z téměř tří desítek smluv upravujících vztahy mezi oběma subjekty celního soustátí {bridging, typ=SUB_SET, na "SR", bridging, typ=SUB_SET, na "český"} *jsou okamžitě vypověditelné všechny...*



Obrázek č. 32: Omezení počtu vztahů asociační anafory

III.5.2.3. Kooperace s TGS – omezení na anotace asociační anafory u závislých uzlů s některými funktory

Při popisu typů asociační anafory jsme již několikrát upozornili na situace, kdy podstata anaforického vztahu je již zahrnuta do syntaktické struktury tektogramatického stromu. V takových případech vztah asociační anafory nezaznamenáváme.

V dané kapitole tyto jednotlivé zmínky zobecníme ve formě pravidel.

Pravidlo 1.

Pokud uzel má funktor `APP`, `MAT`, `AUTH` nebo `PAT`, jeho vztah s přímým rodičem neoznačujeme.

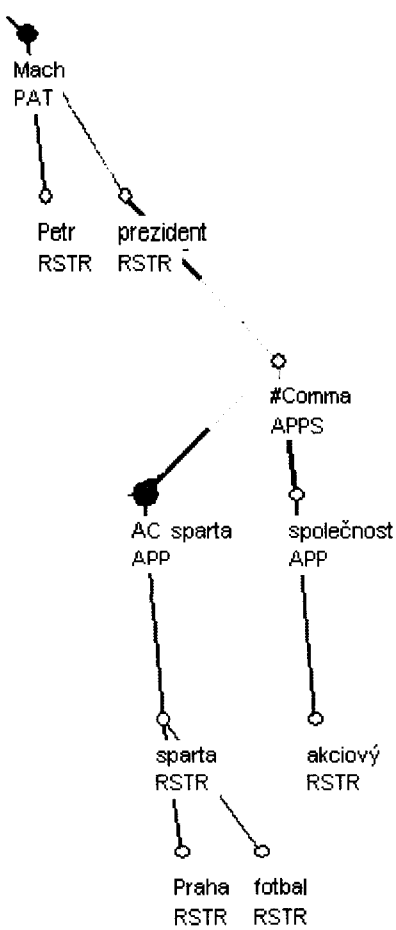
Podle daného pravidla, asociační anaforu neoznačujeme, pokud jsou uzly spojené jednou závislostní šipkou, přičemž závislý uzel má jeden z uvedených funktorů. Takto spojené uzly mají mezi sebou velice často sémantické vztahy, z kterých velkou část je možné zařadit mezi vztahy asociační anafory. Avšak tyto uzly jsou již dostatečně spojené závislostní šipkou a zatěžovat strom dalšími šipkami se jeví v takovém případě zbytečné. Proto např.

nezaznamenáváme asociační anaforu v párech typu *obyvatelka obce* (vztah REST, funktor PAT), *opat kláštera* (vztah FUNCT, funktor APP), *starosta obce* (vztah FUNCT, funktor PAT), *člen výboru parlamentu* (vztah PART, funktor APP), *dílo Wagnera* (funktor AUTH), *jejich obrazy* (funktor AUTH) apod. Srov. také (110):

(110) *Po dílech uvedených autorů* {funktor AUTH, žádný koreferenční vztah} *bude nejspíš vždy slušná poptávka.*

Dané pravidlo funguje také přes spojovací výraz v koordinačních a apozičních konstrukcích, tj. např. vztah FUNCT mezi *prezident* a *akciová společnost AC Sparta Praha – fotbal* použitými v kontextu věty (111) označen také nebude, protože závislé na spojovacím výrazu (v daném případě na čárce) uzly mají funktor APP. Srov. (111) a obrázek č. 33:

(111) *Pokusili jsme se telefonicky kontaktovat prezidenta AC Sparta Praha – fotbal, a. s., Petra Macha, ale jeho vyjádření, zda vyhoví asociaci, jsme nezískali.*



Obrázek č. 33: Kooperace s TGS – neoznačený FUNCT

Podobně jako uvedené funktory, se občas chovají přímé závislosti s funktorem RSTR, srov. např. relativně časté spojení typu *české město*, *maďarský prezident* apod., což je způsobeno tím, že se do koreferenčních řetězců zapojují adjektiva vytvořená od vlastních názvů. V podobných případech asociační anaforu mezi závislým a řídícím uzlem rovněž nezaznamenáváme.

Asociační anaforu však anotujeme v případě, že přímí potomci jsou spojeny jinými funktory, např. DIR1 (viz případ *jeden z faktorů* v III.6.1.)

Pravidlo 2.

Asociační anaforu typu PART neoznačujeme, pokud jsou uzly propojeny přímou závislostí, přičemž závislé uzly mají funktor ACMP.

Srov. *kaplička – daty narození, vlády a smrti; kříž; mramorová pamětní deska* v (112):

(112) *Na břehu Starnberského jezera u místa utonutí byla postavena kaplička s královými daty* {žádný koreferenční vztah na „kaplička"} *narození, vlády a smrti, s křížem* {žádný koreferenční vztah na „kaplička"} *a mramorovou pamětní deskou* {žádný koreferenční vztah na „kaplička"}.

Pravidlo 3.

Asociační anaforu typu CONTRAST neoznačujeme, pokud výrazy, které jsou ve vztahu sémantického nebo kontextového protikladu, jsou přímí nebo nepřímí potomci uzlu s funktory ADVS a CONFR.

Příklady a vysvětlení viz v III.5.1.4.

III.5.3. Nejednoznačný výběr antecedentů

Podobně jako u anotace identické koreference, při výběru antecedentů v anotaci asociační anafory se setkáme s ambiguitními případy. Pro některé typické situace jsme vypracovali anotační strategie. V následující kapitole rozebereme pravidla výběru antecedentu v koordinačních (III.5.3.2.) a apozičních (III.5.3.1.) konstrukcích a v konstrukcích s „kontejnerem" (III.5.3.3.).

III.5.3.1. K otázce výběru antecedentu v případě apoziční konstrukce

V apozičních konstrukcích (anaforu nebo antecedentu), podobně jako u textové koreference (viz B.2.4.1.), šipku vztahu asociační anafory vedeme na konektor (resp. od něj). Srov. např. asociační anaforu typu PART v páru *Západní Německo – městečko za Schirndingem, snad Marktredwitz* v (113)a–b, kde šipka asociační anafory typu PART vede od čárky s funktorem APPS na řídící uzel podstromu *Západni Německo*:

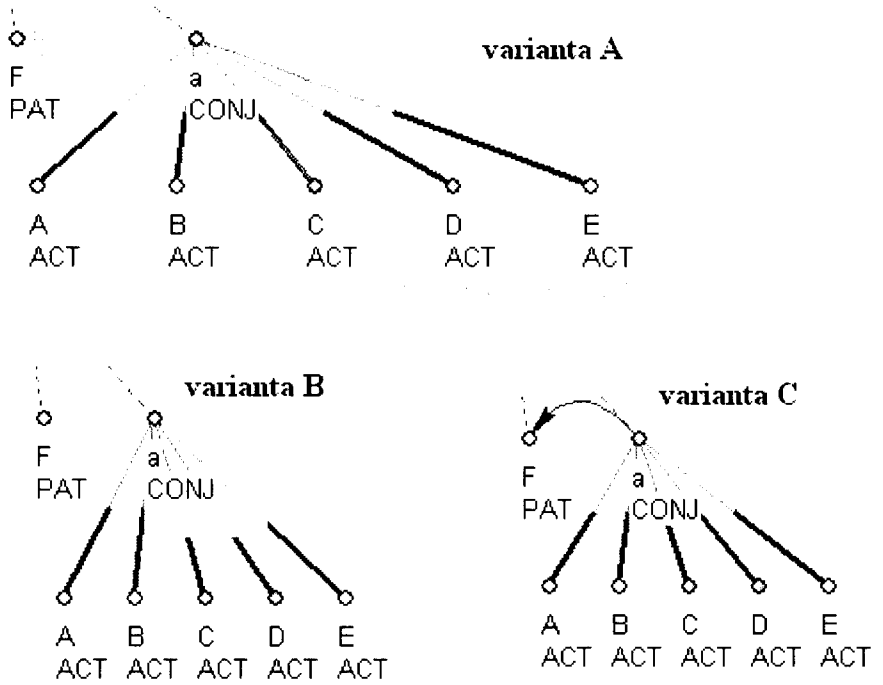
(113) *a. Vzpomínám na takzvané zelené hranice zcela bezbariérové a na dosud nevídanou blahovůli zahraničních a našich celních a policejních orgánů už na jaře 1968, tedy ještě chvíli před záborem, kdy jsme jeli autem na výlet do Západního Německa s Hankou Bělohradskou a Dušanem Hamšíkem.*

b. V nějakém městečku za Schirndingem ,#Comma/APPS {bridging, typ=WHOLE_PART na „Německo"} *snad v Marktredwitzu, jsem si koupil vynikající umělé květiny, jaké se u nás neviděly, také pětatřiceticentimetrovou opici, huňatou, milou a pár podobných i komických drobností jiných.*

III.5.3.2. K otázce výběru antecedentu v případě koordinační skupiny

V případě asociační anafory, kde jeden člen vztahu je koordinační konstrukcí, podobně jako v případě s apozicí (III.4.2.4.1.) a koordinací s textovou koreferencí (III.4.2.4.2.) preferujeme odkaz na formálně řídící uzel s funktorem CONJ (varianta B na obrázku č. 34).

K danému rozhodnutí jsme došli na základě experimentálních pokusů s anotací „podle smyslu". Snažili jsme se neřídit pravidlem „vést na spojku", jak to bylo od začátku stanoveno pro textovou koreferenci a preferovat vztahy mezi autosémantickými uzly (varianta A na obrázku č. 34). Ukázalo se to však jako zbytečné. Jediná výhoda, kterou taková „autosémantická" anotace přináší, je možnost jednoduše vytáhnout páry výrazů ve vztahu asociační anafory, jejich tzv. explicitní viditelnost. To však je možné i při anotaci na spojky při vytažení celých podstromů. Nevýhodou anotace na lexikálně vyjádřené autosémantické uzly je především množení počtu šipek v označení asociační anafory v koordinačních konstrukcích (varianta A na obrázku č. 34).



Obrázek č. 34: Anotace asociační anafory s koordinační konstrukcí

Další nevýhodou je náročná diferenciace mezi asociační anaforou typu SUBSET a textovou koreferencí v případě anotace na autosémantické uzly, závislé na uzlu s funktorem CONJ (varianta C na obrázku č. 34). Předpokládáme totiž, že anotujeme-li vztah mezi autosémantickými uzly v (114)a–b, budeme mít tendenci označovat asociační anaforu typu SUBSET i v (115)a–b, ačkoliv tam jde zřejmě o textovou koreferenci.

(114) *a. Minulé století je bohaté na slavná jména.*
b. Snad vůbec nejvzácnější jsou obrazy Karla Purkyněho a {bridging na "jména" v *a.*} *Jaroslava Čermáka.*

(115) *a. Nedělní vystoupení José Carrerase a Montserrat Caballéové v pražském Rudolfinu (na rozdíl od večera Plácida Dominga ve Sportovní hale) tuto potřebu skutečně naplnilo.*
b. Hudba zněla živě, bez elektroakustického zprostředkování a velikost sálu zároveň poskytovala možnost uspokojivého zrakového kontaktu s umělci {coref_text na spojku "a" v *a.*}.

Vzhledem k uvedeným důvodům asociační anaforu s koordinačními konstrukcemi anotujeme na spojku. Srov. následující příklady (116)–(118):

(116) *a. Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 mil. USD, tj. téměř 2% HDP.*
b. I když letos a {bridging, typ=CONTRAST od spojky „a" na „vloni" v *a.*} *příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu, prognózujeme, že saldo přesto zůstane kladné ve výši 300–600 mil. USD ročně.*

(117) *a. Na toto telefonní číslo však mohou samozřejmě zavolat všichni kluci a děvčata, kteří se ocitnou ve svízelné situaci.*
b. Ptali jsme se několika {bridging, typ=SET_SUB, na spojku „a" v NP (kluci a děvčata) v *a.*}, *jestli by takového kamaráda po telefonu považovali za dobrou věc.*

(118) *Inovovali jsme také receptury pracích prášků, zvýšili podíl účinných látek a parfémů* {bridging, typ=SET_SUB od spojky „a" na „prášek"}.

III.5.3.3. Spojení se slovy s funkcí „kontejneru"

Při anotaci asociační anafory u konstrukcí se slovy ve funkci „kontejneru" (*spousta, řada, milion* apod.) se držíme stejného principu jako při anotaci textové koreference, tj. pokud podle smyslu můžeme odkázat jak na kontejner, tak na jeho závislý uzel, šipku asociační anafory povedeme na „kontejner", abychom dodrželi spojitost textu.

Srov. v následujícím příkladě (119)a–b dva příklady asociační anafory na kontejner: *řada milionářů – Adams* v (119)a a *žvýkačkový – většina gumy* v (119)a–b:

(119) *a. Když o deset let později obrátil ke gumě pozornost louisvilleský lékárník John Colgan, existovala již řada žvýkačkových milionářů (mezi nimi Adams* {bridging, typ=SET_SUB na „řada", nikoliv na „milionář"}*).*
b. Přesto však byly dveře pro zlepšovatele otevřeny dokořán, většina {bridging, typ=SET_SUB na „žvýkačkový" v *a.*} *gumy byla stále ještě jen povrchově oslazený či ochucený kousek chicle.*

(120) *a. Křesťané se modlili za usmíření národů...*

b. Více než tisícový zástup {bridging, typ=SET_SUB na „křesťan" v *a.*} *křesťanů z různých sborů a církví českých zemí a delegace* {bridging, typ= SET_SUB na „křesťan" v *a.*} *křesťanů* {coref_text, typ=NR na „křesťan" v *a.*, funktor APP} *z Německa se v sobotu na vrchu Radobýl u Litoměřic modlil za smíření mezi Čechy a sudetskými Němci.*

Avšak v případech, kdy koreferencí jsou propojeny nejenom kontejnery, ale i jejich závislé uzly, označujeme všechny vztahy – jak kontejnerů tak i to, co obsahují. V párech jako *sklenice vína, krabice gumy* syntakticky závislý uzel bude mít často generickou interpretaci (viz příklady v III.4.2.3.3.).

III.6. Textová koreference nebo asociační anafora. Problematické případy

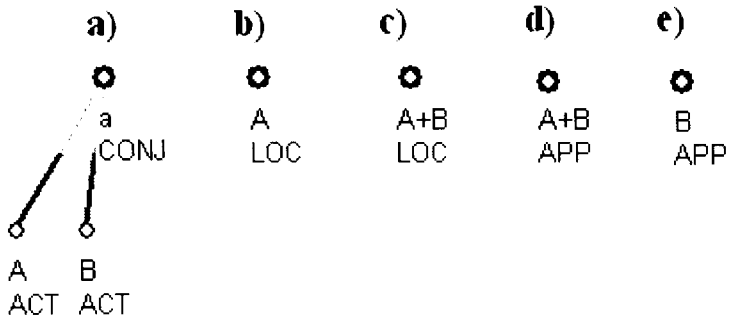
V oddílech III.4 a III.5 jsme vymezili veškeré typy vztahů textové koreference a asociační anafory, které podléhají anotaci na tektogramatické rovině v PDT. Stanovili jsme rovněž některá arbitrární řešení jednotlivých typů. V této části práce se chceme věnovat některým specifickým případům, ve kterých se anotace textové koreference a asociační anafory částečně prolínají.

III.6.1. Dlouhé vzájemně propojené řetězce s textovou koreferencí, asociační anaforou a koordinačními konstrukcemi

Podívejme se na podtržené výrazy v následujícím příkladě (1)a–e:

(1) *a. Malostranské veřejnosti se však nápad poslanců příliš nelíbil: [...] paláce Šternberský a Smiřických i oba domy směrem do Tomášské ulice se staly sídly úřadů a normálními činžáky obývanými nájemníky.*
b. První patro Šternberského paláce skýtá ovšem přeci jen jednu výhodu: terasu vhodnou ke slunění (eventuálně k politickému řečnění), protože je obrácená k jihu na Malostranské náměstí.
c. Kromě pracoven bude v palácových patrech několik kuloárových chodeb... [...]
d. Pražský magistrát pronajme soukromníkům obchody v podloubích obou šlechtických paláců, která směřují do Malostranského náměstí. [...]
e. V nádvoří paláce Smiřických by měla být zřízena dokonce kavárna pro veřejnost, kde by vzhledem k nižšímu podnikatelskému nájmu mohly být i nižší ceny než v okolí, tedy káva za búra místo za třicet.

Jde o situaci, kdy se entity se specifickou referencí vyskytují jako samostatné jednotky (*Šternberský palác* v (1)b, *palác Smiřických* v (1)e), jako koordinační konstrukce (*paláce Šternberský a Smiřických* v (1)a) a jako množina označena jedním výrazem (*paláce, obou šlechtických paláců* v (1)c,d). V takovém případě je otázkou, které vztahy máme zaznamenávat. Schematicky situaci v (1)a–e zobrazuje obrázek č. 35, kde A = *Šternberský palác*, B = *palác Smiřických*:



Obrázek č. 35: Propojené koreferenční, bridging a koordinační vztahy

Spojku „a" v a), [A+B] v c) a d) propojíme textovou koreferencí. Základní otázky, které vznikají v dané situaci, je, kterým vztahem a na které antecedent je třeba anotovat B v e) a A v b).

Pro A v b) jsou možné následující varianty:

- odkázat identickou textovou koreferencí na souřadný uzel A v a);
- odkázat asociační anaforou typu SET_SUB na spojku.

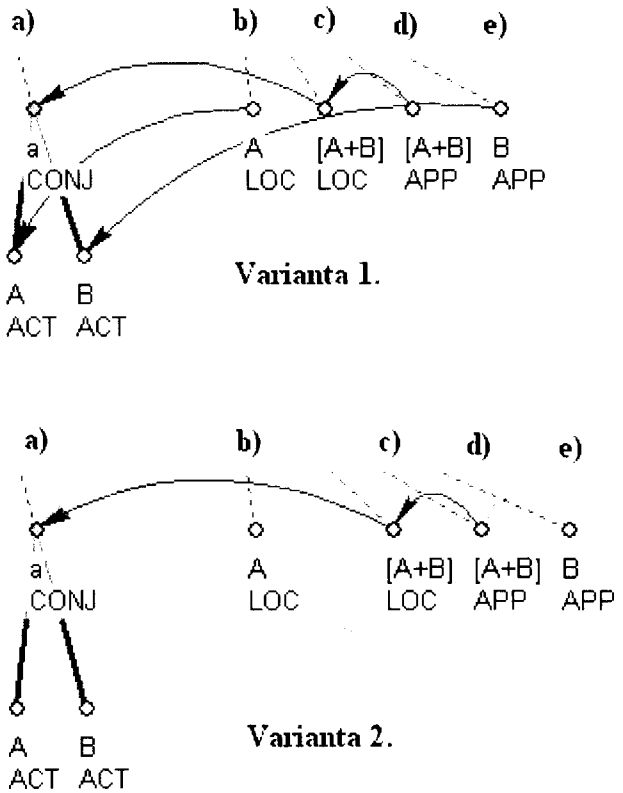
V daném případě pro A v b) je přijatelnější první varianta (na základě principu preference textové koreference a jednotnosti anotace).

Pro B v e) jsou možné následující varianty:

- odkázat identickou textovou koreferencí na souřadný uzel B v a);
- odkázat asociační anaforou typu SET_SUB na [A+B] v d).

V daném případě rozhodnutí již není tak jednoznačné. Odkážeme-li B v e) na B v a), dodržíme princip preference textové koreference, ale ztratíme pro koherenci textu důležitou návaznost mezi B v e) a [A+B] v c) a d), které jsou v kontextu těsně předcházející.

Zachováme-li jednotnost anotace koreference u B v e) a A v b), jsou dvě možné interpretace (srov. obrázek č. 36):

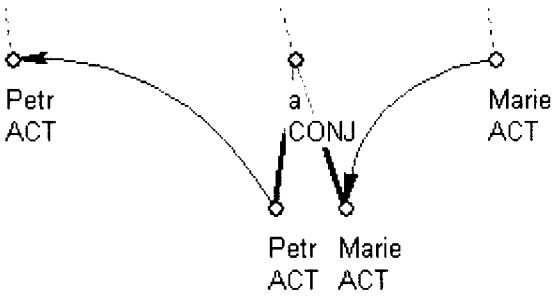


Obrázek č. 36: Propojené koreferenční, bridging a koordinační vztahy

Naše rozhodnutí pro daný případ však neřeší všechny případy problému s propojením jednotlivých entit, koordinačních konstrukcí a množin označených jedním výrazem. Jakmile se setkáme s jinou posloupností uzlů, jsme nuceni znovu hledat příslušné řešení. Srov. např. některé možné kombinace NP *[Petr a Marie], Marie, Petr, vlastníci*:

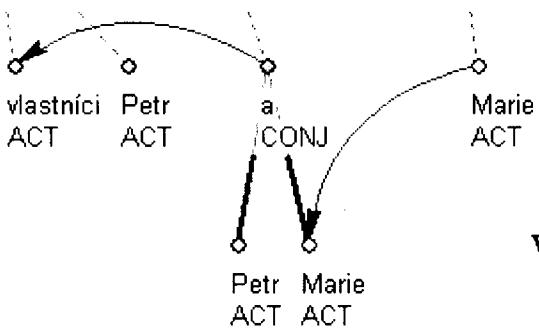
(2) *vlastníci – Petr – Petr a Marie – Marie*

(3) *Petr – Petr a Marie – Marie* (obrázek č. 37)

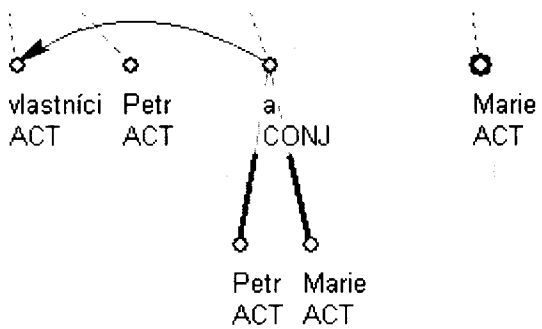


Obrázek č. 37: Propojené koreferenční, bridging a koordinační vztahy

Pokud *Marii* v (2) spojíme textovou koreferencí s *Marii* (obrázek č. 38, varianta 1.), ztratíme informaci o tom, že Marie je vlastníkem. Pokud však *Marii* v (2)a spojíme asociační anaforou typu SUBSET s *[Petr a Marie]* (obrázek č. 38, varianta 2.), anotace nebude jednotná s (3), kde se zobecňující pojem nevykytuje, tedy logicky odkážeme Marii jenom na Marii identickou textovou koreferencí. Poslední možností je označit oba vztahy – textovou koreferenci a asociační anaforu (obrázek č. 38, varianta 3.), nedostatkem tohoto řešení je množení šipek a nejednotnost s (3).

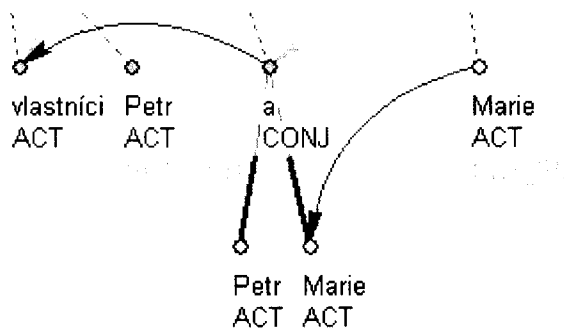


Varianta 1. *Marie - Marie*



Varianta 2. *Marie - [Petr a Marie]*

Varianta 3. *Marie - Marie, Marie - [Petr a Marie]*



Obrázek č. 38: Propojené koreferenční, bridging a koordinační vztahy

Takových variant je potenciálně nekonečná množina a každé řešení má svoje výhody a nevýhody. Pro anotaci koreference a asociační anafory v podobných případech jsme se rozhodli preferovat textovou koreferenci a jednotnost anotace na úkor propojenosti skutečně sémanticky souvisejících uzlů. Vypracovali jsme následující **pravidlo anotace textové koreference a asociační anafory v propojených řetězcích s textovou koreferencí, asociační anaforou a koordinačními konstrukcemi**:

Pokud se v textu kombinují pojmenování jednotlivých entit, množiny těchto entit, označené jedním výrazem a koordinační konstrukce postupujeme následujícím způsobem:

1. Koordinační konstrukce (formálně uzel s funktorem CONJ) propojíme textovou koreferencí s uzly odkazujícími na množinu.
2. Všechny koreferenční výrazy spojíme textovou koreferencí, i když jsou od sebe vzdálenější než uzly, které jsou s anaforickým výrazem ve vztahu asociační anafory.
3. Uzly, které nejsou zapojené do vytvořeného koreferenčního řetězce, připojíme k němu pomocí asociační anafory typu SUBSET.
4. Uzly, které jsou součástí koreferenčního řetězce, ale mají v těsně předcházejícím kontextu antecedentu, se kterým jsou sémanticky spojeny zřetelným vztahem asociační anafory podílející se na kohezi textu, mohou být podle přání anotátora a kontextu propojeny s tímto antecedentem vztahem asociační anafory.

Pro (2) tedy anotujeme variantu 1., případně 3. Pro (1)a–e je správná varianta 1., případně kombinace 1. a 2.

III.6.2. Specifická konstrukce – typ „faktory – jeden z faktorů"

Konstrukce *X – jeden (některé, většina* apod.*) z X-ů* z hlediska anotace koreference a asociační anafory není zcela typická. V tektogramatickém stromě je funktor výrazu *X* v anaforické pozici konvenčně označen jako DIR1. DIR1 je však v prototypické situaci funktorem pro volné doplnění vyjadřující určení místa, které označuje výchozí bod, od kterého směřuje děj vyjádřený řídícím výrazem. Z anotace koreference přímých závislých uzlů ho nemůžeme vyloučit, jak jsme to udělali pro uzly s funktory APP, AUTH aj. (viz III.5.2.3.), protože informace o vztahu mezi danými výrazy by byla v tomto případě ztracena.

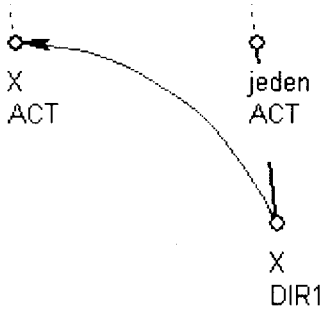
Z těchto důvodů jsme formulovali následující **konvenci označování vztahů mezi jmennými frázemi v konstrukci typu *X – jeden z Xů*** :

V konstrukcích typu *X – jeden (některé, většina* apod.*) z X-ů* postupujeme následujícím způsobem:

1. Jmenné fráze s řídícím uzlem *X* propojíme textovou koreferencí typu 0 nebo NR podle smyslu.
2. Výraz *jeden* propojíme vztahem asociační anafory typu SET_SUB s druhým *X*-em.

Toto řešení se poněkud vymyká z doposud formulovaných pravidel a principů. Považujeme však tuto konstrukci za speciální a její řešení za výjimku.

Graficky anotace konstrukce typu *X – jeden z Xů* vypadá následujícím způsobem (obrázek č. 39):



Obrázek č. 39: Schéma anotace konstrukce "X – jeden z X-ů"

Srov. následující příklady:

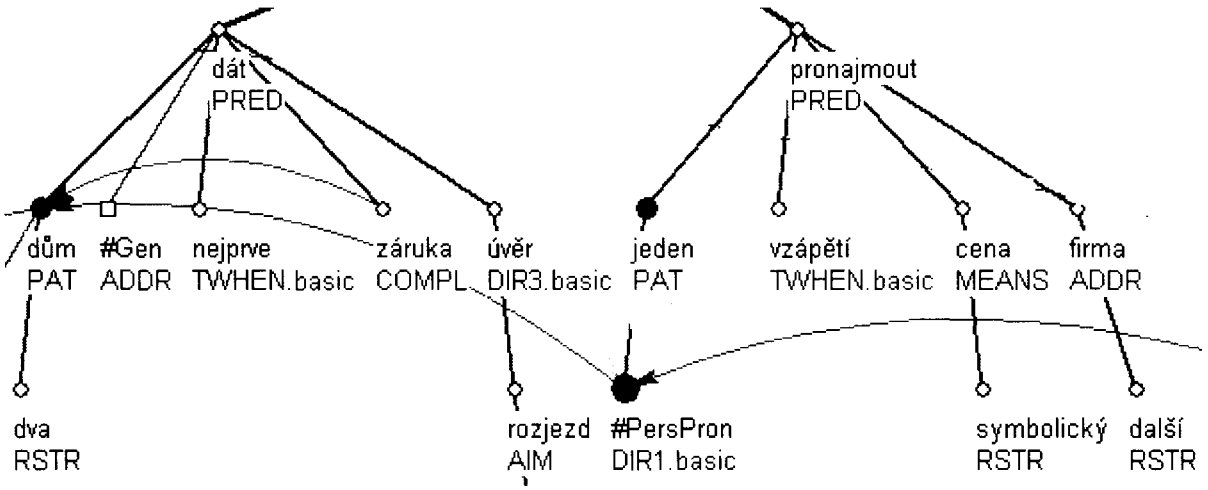
(4) *a. V praxi se tato rovnováha realizuje tím, že se každý faktor*[115] *chová rutinním a adaptivním způsobem.*
b. Dynamika kapitalistického systému vzniká teprve tím, že se jeden {bridging, typ=SET_SUB na „faktor" v *b.*} *z faktorů* {coref_text, typ=0 na „faktor" v *a.*} *začne chovat netradičně.*

Srov. také (5) a obrázek č. 40 s původní pronominální koreferencí:

(5) *Tyto dva domy nejprve správce SFK Ladislav Kratochvíl, který je zároveň náměstkem ministra kultury, dal coby záruku na úvěr pro rozjezd České lotynky a jeden* {bridging, typ=SET_SUB na „#PersPron"} *z nich* {coref_text, typ=0 na

[115] V daném případě sice jde o distributivní referenci, avšak v kontextu to sémantický odpovídá výrazu „všechny faktory", proto ji považujeme za množinu se specifickou referencí.

„dům"} *vzápětí za symbolickou cenu pronajal další firmě.*



Obrázek č. 40: Specifická konstrukce – typ „X – jeden z X-ů"

III.6.3. Případ „zaměstnanci – každý ze zaměstnanců"

Konstrukce *X – každý z X-ů* je syntakticky velice podobná konstrukci *X – jeden (některé, většina* apod.*) z X-ů* (III.6.2.), má však zcela odlišnou referenční hodnotu. Podobně jako *jeden* v *jeden z X-ů*, výraz *každý* má v *každý z X-ů* substantivní platnost, avšak zatímco *jeden* v *jeden z X-ů* referuje na podmnožinu referentů závislého na něm uzlu, referenční potenciál *každý* v konstrukci *každý z X-ů* je totožný s X. Proto považujeme za smysluplné anotovat konstrukci *každý z X-ů* jako jeden celek (podobně jako bychom anotovali případ NP *každý X,* kde *každý* by již byl formálně závislý na *X*), přičemž za koreferující uzel konstrukce považujeme její závislý uzel *X*. Řídící uzel *každý* ponecháváme nepropojený. Srov. např. (6)a–c, (7)a–b a (8)a–c a obrázek č. 41:

(6) *a. Podle přesvědčení majitelů dosáhla prosperity zejména proto , že zaměstnává lidi, na které* { coref_gram, na „člověk"} *se může spolehnout.*
b. Kritéria výběru jsou přísná.
c. Každý ze zaměstnanců {coref_text, typ=NR od „zaměstnanec" na „který" v *a.*} *musí být odborníkem.*

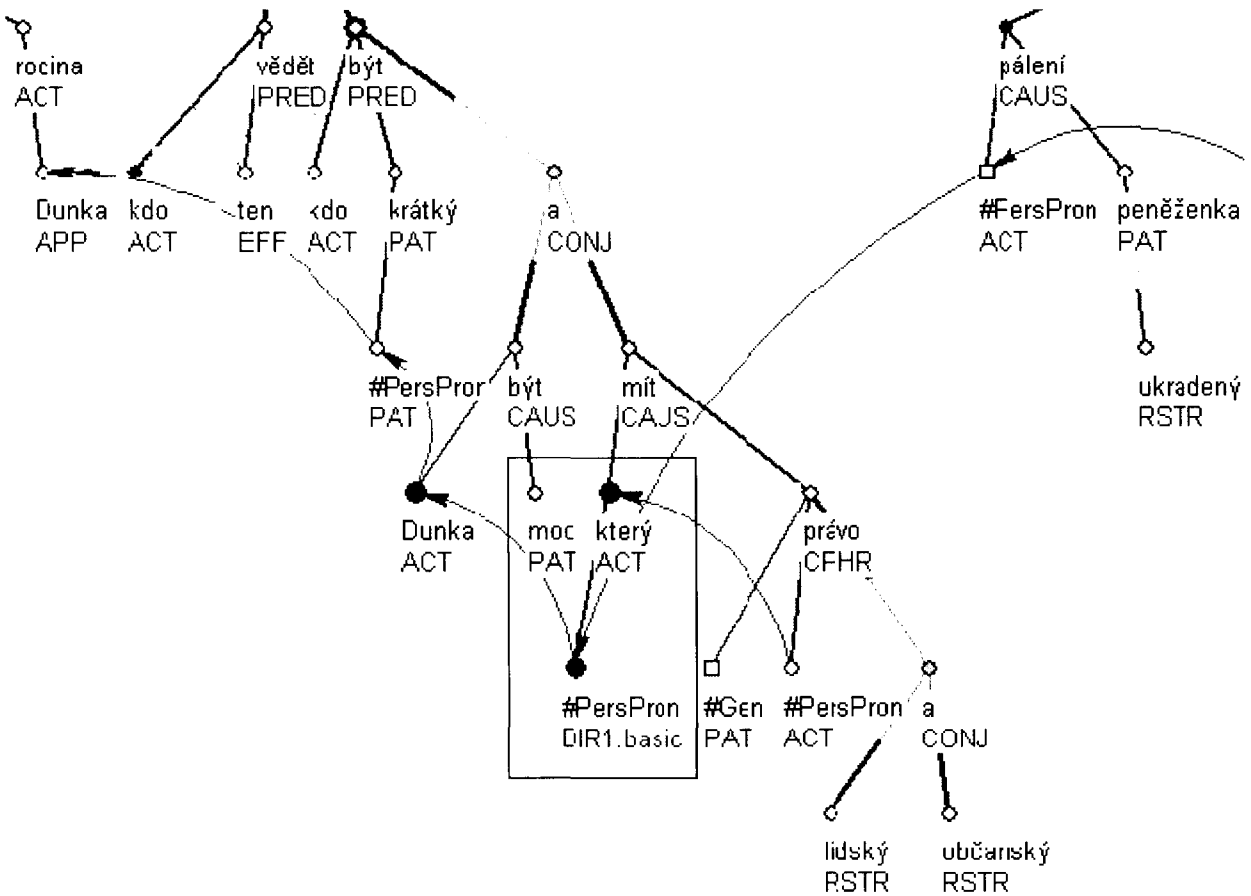
(7) *a. Celkem šest koncertů obsahuje čtvrtfinále soutěže rockových skupin Marlboro Rock-in '95, které začalo v pátek v trutnovském klubu Eden.*
b. Na každém z koncertů {coref_text, typ=0 od „koncert" na „koncert" v *a.*}

vystoupí čtyři ze dvaceti čtyř skupin, které prošly prvním výběrovým sítem.

(8) *a. Portmonky měla v referátu rodina Dunkových.*

b. Všichni to věděli a všichni na ně byli krátcí, protože Dunků bylo moc a každý z nich měl svá lidská a občanská práva.

c. Z pálení ukradených peněženek se kolem jejich chalupy linul penetrantní čmoud.



Obrázek č. 41: Specifická konstrukce „X – každý z X"

III.6.4. Propojení koreferenčních řetězců jediným vztahem asociační anafory

V případě, že dva koreferenční řetězce jsou mezi sebou ve vztahu asociační anafory, šipku asociační anafory vedeme pouze jednou, a to při prvním výskytu daného vztahu asociační anafory. Dále tyto řetězce považujeme za propojené a asociační anaforu mezi dalšími elementy těchto řetězců již nezaznamenáváme.

Srov. propojení koreferenčních řetězců odkazující na *Alžír* a *Alžířany* v (9)a–h:

(9) *a. Alžírští extremisté vyhrožují Maroku.*

b. Alžír {coref_text, typ=0 na „alžírský" v *a.*} –

c. Islámští fundamentalisté z Alžírska {coref_text, typ=0 na „Alžír" v *b.*} *včera zveřejnili výhrůžku na adresu Maroka, že sáhnou k násilí, budou-li Alžířané* {bridging, typ=REST na „Alžírsko" v *c.*} *v Marockém království vystaveni špatnému zacházení.*

d. Výhrůžku vydala Islámská fronta spásy (FIS), jejíž zákaz po volbách v roce 1992 vyvolal v Alžírsku {coref_text, typ=0 na „Alžírsko" v *c.*} *katastrofální vlnu násilí a teroru.*

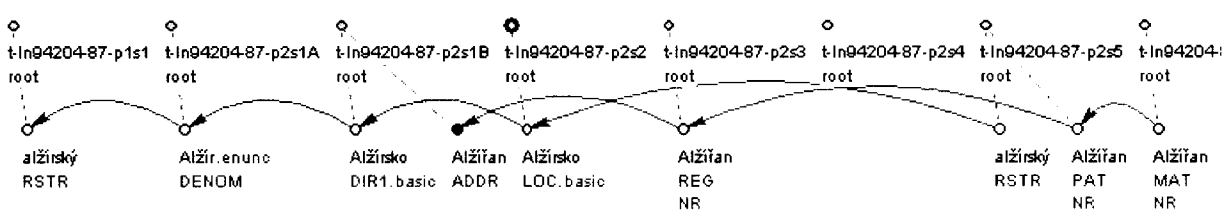
e. Maroko včera poněkud zmírnilo svou vízovou politiku vůči Alžířanům {coref_text, typ=0 na „Alžířan" v *d.*}, *když v několika marockých městech zazněly hrozby pumových atentátů.*

f. Kvůli falešné pumové hrozbě v Marrákeši stovky lidí opustily správní budovu v centru města a podobně se rychle vylidnilo železniční nádraží v Rabatu.

g. Alžírští {coref_text, typ=0 na „Alžírsko" v *g.*} *muslimští extremisté uvedli, že marocká policie s Alžířany* {coref_text, typ=0 na „Alžířan" v *e.*} *špatně zachází.*

h. Stovky Alžířanů {coref_text, typ=0 na „Alžířan" v *g.*} *byly údajně zavlečeny na policejní stanice, kde byly vyslýchány.*

Schematicky je možné propojení koreferenčních řetězců jediným vztahem asociační anafory zobrazit následujícím způsobem:



Obrázek č. 42: Propojení koreferenčních řetězců jediným vztahem asociační anafory

III.7. Speciální typy reference (coref_special)

Na tektogramatické rovině PDT jsou samostatně zachyceny některé typy odkazování, kdy antecedentem není konkrétní uzel či podstrom tektogramatického stromu. Jde o případy exoforického odkazování k mimotextové situaci či skutečnostem (hodnota `exoph` atributu coref_special) a odkazování k segmentu textu většímu než jedna věta (hodnota `segm` atributu coref_special). Ve fázi anotace původní pronominální koreference byly tyto speciální typy reference zpracovány pro stejné typy výrazů, jako v anotaci ostatních případů textové koreference (osobní a ukazovací zájmena v 3. osobě v platnosti substantiva, III.2.1.1. a Mikulová a kol. 2005, s. 1043n.) V dané fázi zpracování rozšířené textové koreference a asociační anafory doplňujeme existující vztahy novými případy, kdy jako odkazující element vystupují jiné slovní druhy (viz III.2.1.– III.2.4.).

III.7.1. Exoforické odkazování

Exoforický vztah anotujeme v případě, že jmenná fráze explicitně poukazuje k mimotextové situaci či skutečnostem. V rozšířené anotaci koreference a asociační anafory odkazovací NP musí přitom obsahovat explicitní identifikátor (ukazovací zájmeno).

Můžeme vyčlenit následující skupiny nových exoforických odkazu:

1. **Časová** (*v tomto roce*) **a prostorová** (*na tomto místě*) **deixe**. Srov. např. větu (1) z článku v novinách, kde *v těchto dnech* odkazuje k aktuálnímu času publikace novin:

(1) *Dokončeny by měly být do 31 . prosince 1995 , a to i přes jisté zdržení způsobené opožděným stěhováním nájemníků z domů čp. 8 a 518 do náhradních bytů na sídlišti Barrandov v těchto dnech* {coref_special, typ=`exoph`}.

Srov. také (2)–(4):

(2) *Devadesátiny Velkého architekta připadly na počátek tohoto týdne* {coref_special, typ=`exoph`} *a obešly se bez bombastických oslav.*

(3) *Za prvních sedm měsíců tohoto roku* {coref_special, typ=`exoph`} *dosáhly registrace nových osobních vozů počtu 2017474, což bylo o 0.1 procenta více než ve stejném období loni.*

(4) *Plukovník StB Čadek udělal chybu, když poslechl doporučení novinářů a uchýlil se do předem připravených pozic. Neměl podlehnout panice a v klidu vyčkat věcí příštích. Vždyť v téhle zemi* {coref_special, typ=exoph} *se zatím žádnému skutečně velkému podvodníkovi nezkřivil ani vlas na hlavě.*

2. **Deixe u pronominálních adverbií**, u kterých nově anotujeme rozšířenou textovou koreferenci

(5) *A tu* {coref_special, typ=exoph} *se dostáváme zpět k počátku tohoto textu .*

3. **Exoforický odkaz na celý text** jako na objekt:

(6) *Informace v tomto přehledu* {coref_special, typ=exoph} *jsou bezplatnou službou podnikatelům.*

Odkazy typu exoph dodáváme pouze v případě opravdové exoforické deixe (je možné představit, že mluvčí při promluvě ukazuje prstem).[116] Exoforický odkaz **nezaznamenáváme** u jiných mimotextových odkazů, například:

- kdy deiktický výraz je součástí lexikální sémantiky daného slova (výrazy typu *dnes, zítra, letos, současnost* apod.),
- v jiných konstrukcích se „šifterovou" sémantikou typu *příští rok, v současné době, minulý týden, poslední rok, letošní rok, současná situace, v sobotu, v červenci* apod.
- odkazy na generalizovanou množinu čtenářů v první osobě:

(7) *a. Zákon o prostituci se u nás teprve připravuje.*
b. Většina z nás {žádný koreferenční odkaz} *má hodně cholesterolu.* (jako název článku)

- exoforické odkazy na příznak:

(8) *Angel říká, že fronty se každým dnem znatelně prodlužují. „Viděl jsem šňupat opravdový dámy, jsou tu i lidi, který vypadaj, jako by umírali na AIDS. Je hrozný, jak jim takovýhle život* {žádný koreferenční odkaz} *užírá rozumný myšlení rychleji*

[116] Kromě případů časové deixe.

než blesk."

V případě potřeby některé z těchto odkazů se dají dohledat automaticky.

III.7.2. Odkazy na segmenty textu

Odkazy na segmenty textu dodáváme v následujících případech:

1. Jmenná fráze (obvykle s identifikátorem) odkazuje na více než jednu větu v předchozím kontextu. Jsou to především anaforické jmenné fráze typu *tento pohled, tento problém, tento směr, tento přístup* apod. Srov. např. *v tomto směru* v (9)a–c odkazující na dvě věty (9)a–b:

(9) *a. Celní unie bude sice existovat na papíře ještě dalších dvanáct měsíců, ale v praxi dostanou vzájemné vztahy punc tvrdosti mezinárodního obchodu.*
b. Poroste administrativa.
c. Jistotu v tomto směru {coref_special, typ=segm} *dávají nejnovější kroky vlády SR, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience.*

Srov. také odkaz na větší a méně zřetelný antecedentní kontext:

(10) *a. V článku jsme odpovídali na dotaz naší pardubické čtenářky, kde by měla uzavřít životní pojištění, aby platila co nejméně a získala co nejvíce. Při výběru pojišťovny jsme zvažovali, kolik by musela zaplatit ročně na pojistném, zda by se mohla připojistit na úraz, zda by byla okamžitě po uzavření pojistné smlouvy pojištěna na sjednanou pojistnou částku a konečně zda si bude moci v případě náhlé potřeby vypůjčit větší sumu peněz z dosud zaplaceného pojistného, aniž by to mělo vliv na výši pojistné částky. Čtenářce jsme doporučili pojistit se u První americko-české pojišťovny, u které jsme zjistili druhé nejlevnější pojistné, možnost úrazového připojištění a výběru peněz prostřednictvím půjčky a pojištěna by byla okamžitě z titulu životního pojištění i úrazového připojištění na sjednané pojistné částky. Do redakce nám přišly ohlasy. Z dopisu Bedřicha Kováře, ředitele úseku pojištění osob v České pojišťovně vyjímáme: Podíl na zisku je v článku zmíněn, avšak v případě pojištění nabízeného Českou pojišťovnou je důležitou skutečností, že pojištěný má mimo sjednanou pojistnou částku zaručenou zvláštní prémii a*

navíc valorizaci.

b. V případě dožití se podle uvedeného příkladu {coref_special, typ=segm} *by naše hypotetická pojištěná měla zaručen nárok na zvláštní prémii ve výši 25%. V případě smrti je rovněž zaručena zvláštní prémie, jejíž výše je závislá na délce trvání pojištění.*

b. V uvedeném příkladu {coref_special, typ=segm} *by byla minimálně 30% a maximálně 50%.*

2. Jako coref_special, typ=segm řešíme také případy odkazů na technicky nevyčlenitelné podstromy uvnitř jednoho stromu (viz III.4.2.3.5.1.)

3. Odkaz na segment nemusí být s antecedentním segmentem textu ve vztahu identické koreference. I když vzácně, vyskytují se také odkazy na segmenty textu, které jsou s ním ve vztahu asociační anafory. Jde o kontexty, kdy se např. po vyprávění o jednom problému, následuje věta typu *Podobné problémy řešíme také v naši oblasti.*

Podobná situace nastává v kontextech s implicitním odkazem na dobu událostí, popisovaných v předcházejícím kontextu, přičemž kontext popisující událost je větší než jedna věta (jinak viz podtyp ANAF, III.5.1.5.). Na dané etapě anotace takové případy označujeme jako segm. Srov. (11)–(13):

(11) *a. Tak jako každý Mexičan, i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle.*

b. Tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku.

c. Právě v té době (kdy se to všechno dělo – *A.N.*) {coref_special, typ=segm} *přihrála náhoda Santa Annovi do cesty Thomase Adamse, fotografa a především vynálezce všeho druhu.*

(12) *a. Koho chce Bůh potrestat, toho raní slepotou, říkávala moje babička. Ostatně není to právě přehlížívý přístup některých představitelů ODS ke všem oponentům, k novinářům, k veřejnému mínění, co způsobilo setrvalý pokles volebních preferencí této strany? Zbývají dva roky a procent stále ubývá. ODS jako by chtěla dokázat, co někteří autoři slovních průjmů již nějaký čas tvrdí: že totiž ona sama je*

momentálně tou jedinou silou, která ji může porazit. Bylo by asi neradostné, kdyby současná vláda prohrála hned příští parlamentní volby.

b. Ne tak kvůli ní samé, jako kvůli ekonomické i politické transformaci, jíž je garantem a která tou dobou {coref_special, typ=segm} *ještě nebude zcela u konce.*

(13) *a. Viděl ji před očima, jak stojí přitištěna ke zdi Sabinina ateliéru a vráží si jehly pod nehty. Vzal do rukou její prsty, hladil je, dal si je ke rtům a líbal je, jako by na nich byly ještě stopy krve.*

b. Ale od té doby {coref_special, typ=segm} *jako by se proti němu všechno spiklo. (SYN2005)*

Občas můžeme potřebovat (nikoliv realizovat) kataforický odkaz segm dopředu. Následující příklad v anotaci nezaznamenáváme, pouze ho zde uvádíme jako jednu z možných forem:

(14) *a. Do redakce nám přišly ohlasy* (tady by se hodil segm dopředu).

b. Z dopisu Bedřicha Kováře, ředitele úseku pojištění osob v České pojišťovně vyjímáme:

c. Podíl na zisku je v článku zmíněn, avšak v případě pojištění nabízeného Českou pojišťovnou je důležitou skutečností, že pojištěný má mimo sjednanou pojistnou částku zaručenou zvláštní prémii a navíc valorizaci.

III.7.2.1. Hraniční případy mezi typem coref_special, typ=segm a asociační anaforou typu SUBSET

V některých případech odkaz na segment textu je konkurenční s asociační anaforou typu SUBSET. Srov. vztah mezi (15)a–b a *to* v (15)c:

(15) *a. Spolupráce by měla dostat patřičný rytmus, režim.*

b. Vysoké využití podhorských pastvin , nejkvalitnější stáda.

c. To jsou předpoklady pro výrobu kvalitních potravin.

V původní anotaci pronominální koreference ukazovací zájmeno *to* odkazuje na širší předchozí kontext jako coref_special, typ=segm. Taková anotace je správná, chápeme-li ukazovací zájmeno *to* jako odkaz k celým dvěma větám (15)a a (15)b, tj.

to = spolupráce má dostat patřičný rytmus, režim, vysoké využití podhorských pastvin, nejkvalitnější stáda.

To však není jedinou možnou interpretací. Ukazovací zájmeno *to* může odkazovat pouze k jmenným frázím, nikoliv k predikaci v (15)a, tj. jako

to = předpoklady = {patřičný rytmus, režim, vysoké využití podhorských pastvin, nejkvalitnější stáda, ...}.

Při výběru mezi speciální textovou koreferencí typu segm a asociační anaforou typu SUBSET platí následující **pravidlo preference typu SUBSET**:

> Pokud můžeme spolehlivě odkázat asociační anaforou typu SUBSET, preferujeme SUBSET, nikoliv speciální textovou koreferenci typu segm.

V (15)c tedy tedy preferujeme asociační anaforu typu SUBSET.

Srov. také odkaz na několik slovesných frází v (16)a–h:

(16) *a. - Stála vás kvalita hodně peněz a potu?*

b. - Museli jsme se přizpůsobit tržní filozofii.

c. Dříve jsme měli za úkol jen nasytit trh množstvím výrobků a na jakost se nehledělo.

d. Nyní jsou požadavky opačné.

e. Proto jsme zpřísnili vlastní kontrolu.

f. Inovovali jsme také receptury pracích prášků, zvýšili podíl účinných látek a parfémů.

g. U detergentu Toto jsme například řešili problém s udržením stálé kvality, protože jednotlivé partie byly nevyvážené.

Investovali jsme dva miliony korun do nákupu pásových vah, zpřesnili dávkování a jakost pracího prášku stabilizovali.

h. - V těchto opatřeních {bridging, typ=SUB_SET na „zpřísnit", „inovovat", „zvýšit", „řešit", „investovat", „zpřesnit", „stabilizovat"} *vidíte podstatu komerčního úspěchu?*

III.8. Zásah do anotace původní zájmenné koreference

Na závěr kapitoly o schématu anotace rozšířené textové koreference a asociační anafory na tektogramatické rovině uvedeme několik poznámek k vyrovnání anotace původní pronominální koreference s tím, co bylo ve stávající anotaci nově zavedeno.

Vztahy pronominální textové koreference, které se neúčastní koreferenčních řetězců prodloužených rozšířenou koreference, zůstali většinou beze změn, ani jsme je důsledně nekontrolovali.

V pronominálních koreferenčních řetězcích, které byly prodloužené rozšířenou koreferencí, jsme kontrovali následující:

1. Typ reference – u pronominalizací s generickou koreferencí „defaultní" typ 0 se měnil na typ NR. Srov. (1):

(1) *Dítě je ještě z formy, ale už #PersPron* {coref_text, typ=NR na "dítě"} *musí plnit doma i ve škole spoustu úkolů a působí mu* {coref_text, typ=NR na "#PersPron"} *to problémy.*

Taková změna původní anotace má však dvě problematické stránky. Za prvé, nekontrolujeme důsledně všechny pronominalizace, je tedy možné, že některé vztahy generické povahy zůstanou s typem 0 a zkreslí typologickou statistiku. Za druhé, následná kontrola již oanotovaných vztahů je matoucí a časově náročná. Tento problém je však možné řešit automatickou změnou typu 0 na typ NR, v případě, že je pronominalizace součástí generického koreferenčního řetězce.

2. Dalším možným zásahem do anotace původní zájmenné koreference jsou případy oanotované pronominální koreference, která de facto není koreference, ale jiný typ vztahů, který ve fázi anotace rozšířené textové koreference a asociační anafory již umíme zaznamenat. Srov.:

(2) *V ČR podniká 80 zásilkových firem, nejvíce v Praze, kde jich* {bridging, typ=SUB_SET na "firma"} *působí 35.*

(3) *První nákup realizovala v severních Čechách, postupně ho* {bridging, typ=ANAF na "nákup"} *pořídí ve všech částech ČR.*

(4) *Dovoz ze států ESVO převýšil náš vývoz o 13.9 miliardy korun a* [ACT] {bridging, typ=ANAF na "dovoz"} [PAT] {bridging, typ=ANAF na "vývoz"} *z Evropské unie o 2.1 mld Kč.*

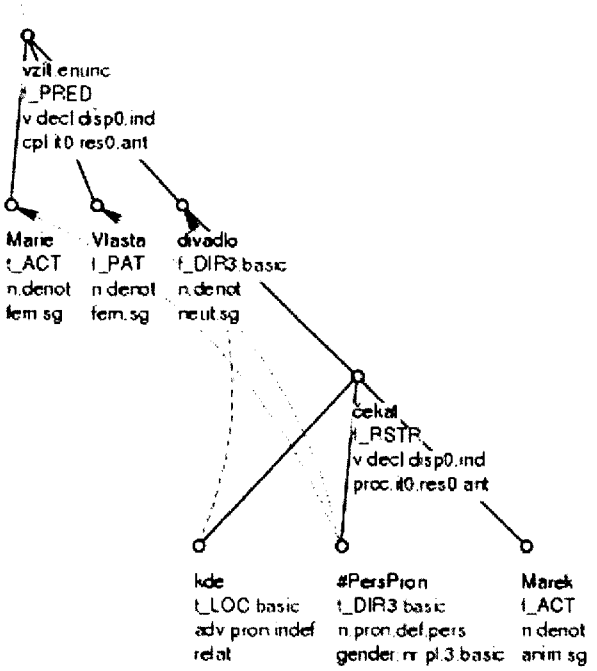
Použití pronominální asociační anafory je jev celkově diskutabilní. Tak např. Erkü a Gundel (1987) tvrdí, že použití asociační anafory není přípustné nebo velice příznakové, zatímco Yule (1979) je naopak zastáncem toho, že její výskyt je zcela běžný. Cornish (2007) rozlišuje dva základní typy implicitních referentů – centrální a periferní – a předkládá hypotézu, že zatímco periferní referenty (které vyvolávají např. způsob děje, instrument apod.) nemohou sloužit jako antecedenty pro pronominální asociační anaforu, centrální referenty jsou jimi zcela běžně. Výsledky dvou praktických experimentů však ukázali, že asociační pronominální anafora na periferní referenty způsobuje pomalejší porozumění textu.[117]

Během anotace rozšířené koreference na PDT pronominální odkazy na nekoreferenční antecedent se vyskytují zcela běžně. Podobné příklady anotujeme jako ANAF, příp. SUBSET. Tak v (2) jsme se rozhodli pro asociační anaforu typu SUBSET, v (3) a (4) – pro nekoreferenční anaforu typu ANAF.

3. Případy, kdy antecedentem textově koreferenčního vztahu jsou dva uzly tektogramatického stromu. Srov. v anotaci pronominální koreference v (5) antecedentem osobního zájmena *na ně* (reprezentovaného v tektogramatickém stromě uzlem s t-lematem `#PersPron`) jsou dva uzly (*Marie, Vlasta*), ke kterým se odkazuje textovou koreferencí jednotlivě. Srov. (5) a obrázek č. 43:

(5) *Marie vzala Vlastu do divadla, kde na ně čekal Marek.*

[117] Viz Cornish 2007, s. 30.



Obrázek č. 43: Odkazování textovou koreferencí k několika antecedentům

Takové řešení v anotaci pronominální koreference však bylo podmíněno absencí jiných vztahů než koreferenčních. V dané fázi anotace, kdy máme k dispozici asociační anaforu typů SUBSET a PART, může z jednoho uzlu vycházet pouze jedna šipka textové koreference. Případy typu (5) měníme na asociační anaforu typu SUBSET. Tedy referenční vztahy v (5) budou vypadat takto:

(6) *Marie vzala Vlastu do divadla, kde na ně* {bridging, typ=SUB_SET na „Vlasta" a na „Marie"} *čekal Marek.*

Srov. také. (7), kde původní pronominální koreference anotována nebyla a ve stávající anotací se označuje jako asociační anafora typu „množina – podmnožina":

(7) *a. [...] paláce Šternberský a Smiřických i oba domy směrem do Tomášské ulice se staly sídly úřadů a normálními činžáky obývanými nájemníky.*
b. Ti {bridging, typ=SUB_SET, na „úřad" a „nájemník" v *a.*} *se nyní měli vystěhovat a někteří občané to pochopili jako znárodňovací akt.*

Dané pravidlo má však jedno důležité omezení. Pokud antecedenty vztahu jsou (technicky) přímí potomci spojky, šipka textové identické koreference vede na tuto spojku (viz III.4.2.4.1. – III.4.2.4.2.). Srov. (8)a–b:

(8) *a. Asi rok se Adams a jeho nejstarší syn snažili – chicle vařili, čistili, přidávali množství různých látek a míchali s pravým kaučukem.*

b. Když asi po roce #PersPron {coref_text, typ=0 na spojku „a" v *a.*, nikoliv bridging, typ=SUB_SET na „Adams" ani „syn"} *své úsilí vzdali, rozhodl se Adams, že vše, co mu z chicle ještě zbylo, hodí do řeky.*

IV. Aplikace a evaluace probíhající anotace

Ačkoliv hlavním zaměřením naší práce je popis teoretického schématu zpracování anafory a koreference, považujeme za vhodné také stručně představit aplikaci našeho schématu na anotaci rozšířené textové koreference a asociační anafory na tektogramatické rovině v PDT a uvést některé její statistické a evaluační výsledky.[118] V této části práce představíme technickou bázi, o níž se opírá naše anotace (viz IV.1.), uvedeme dále statistiku oanotovaných uzlů a vztahů (viz IV.2.), výsledky testů mezianotátorské shody (viz IV.3.) a pokusíme se některé z těchto výsledků vysvětlit (viz IV.4.).

IV.1. Technické provedení

IV.1.1. Formát dat

Anotace je uskutečňována pomocí multiúrovňového anotačního nástroje TrEd (Pajas – Štěpánek 2008) vypracovaného na ÚFALu UK Praha. Pro účely anotace koreference a asociační anafory bylo rozšířeno standardní datové schéma TrEdu – byly doplněny atributy TYPE u textové koreference a zaveden zvláštní atribut pro asociační anaforu (podrobněji viz Nedoluzhko a kol. 2009) Pro anotaci se využívá speciální anotační rozšíření (Bridging Anaphora Extension), které poskytuje anotátorům pohodlný náhled a systém maker (viz IV.1.2.). Typický vzhled TrEdu je zobrazen na obrázku č. 46 v IV.1.2.2.

Pro anotaci koreferenčních a jiných diskurzivních vztahů se v posledních pěti letech nejvíce používají anotační nástroje MMAX-2 (Müller – Stube 2003) a PALinkA (Orăsan 2003). Tyto nástroje jsou přizpůsobeny manuální anotaci koreferenčních a anaforických vztahů a jsou aplikovatelné pro různá anotační schémata. Jsou však prováděny na celých textech, nikoliv na stromech a předpokládají především dvoufázovou anotaci s doplněním velkého počtu příznaků k jmenným frázím (viz II.3.).

IV.1.2. Pomoc anotátorům

Anotace koreferenčních a anaforických vztahů na tektogramatické rovině je manuální. Pro ulehčení náročné ruční práce jsme však vypracovali několik pomocných nástrojů. Tyto nástroje jsou dvojího druhu. Za prvé, používá se předanotace koreferenčních vztahů v párech výrazů, které mají velkou pravděpodobnost být koreferenční (IV.1.2.1.). Za druhé, řada nástrojů se zapojí během anotace samotné (IV.1.2.2.).

[118] První výsledky naše anotace byly již publikovány v (Nedoluzhko 2009, Nedoluzhko a kol. 2009a–c).

IV.1.2.1. Předanotace dat

Předanotace dat se aplikuje v PDT na základě seznamu párů slov, mezi nimiž bude s velkou pravděpodobností koreferenční vztah. Během anotace anotátoři ověřují a podle nutnosti upravují již zaznamenanou koreferenci. Takový přístup se osvědčil jako jednoznačně rychlejší ve srovnání s manuální anotací. Páry výrazů, anotované automaticky jsou následujících typů:

- absolutní většinu tvoří páry *podstatné jméno (NE)*[119] *– odvozené od NE adjektivum*, v různých kombinacích (např. *německý – Německo, pražský – Praha, Praha – Praha, německý – německý* apod.);
- zbývající jsou páry výrazů, kde alespoň jedním členem páru je zkratka odvozená od NE (např. *ČR – Česko*).

Tyto páry byly automaticky vybrány z morfologického parseru pro češtinu. Tento seznam obsahuje více než šest tisíc párů typu *německý – Německo, ČR – Česko, Rusko – ruský* apod. Seznam se průběžně doplňuje a opravuje.

Předanotace se týká pouze textové koreference. U asociační anafory je předanotace teoreticky také možná, ale zatím nebyla implementována.

IV.1.2.2. Automatická pomoc v průběhu anotace

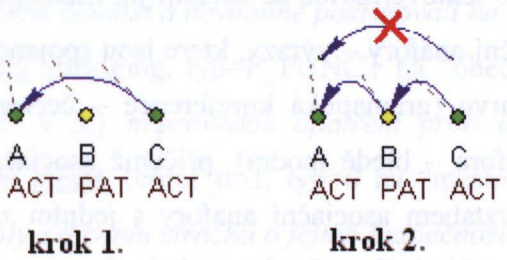
Na anotačním nástroji TrEd byly pro pomoc anotátorům implementovány následující vlastnosti:

1. **Manuální předanotace**. Pokud anotátor zaznamená v textu výraz, který se často opakuje a referuje na tutéž entitu, může jedním kliknutím myší popojit všechny uzly s daným t_lemmatem do jednoho koreferenčního řetězce.

2. **Vyhledávání nejbližšího antecedentu**. Princip dodržování koreferenčního řetězce tvrdí, že jako antecedent koreferenčního vztahu vždy slouží poslední (nejbližší v textu) výraz, koreferenční s daným výrazem (viz III.1.2.). Tento princip se však v některých případech těžko dodržuje manuálně, zvláště při anotaci na textech, nikoli na stromech (posledním předcházejícím uzlem může být např. nevyjádřený uzel s t_lemmaty `#PersPron`, `#Cor`, `#Qcor` aj.). V případě, že anotátor povede koreferenční vztah na jiný, než na nejbližší předcházející uzel, nástroj automaticky přesměruje nově označenou koreferenční šipku na nejbližší koreferující uzel v již existujícím koreferenčním řetězci. Vytvořený koreferenční řetězec však může být dále modifikován. Například, máme-li posloupnost ze tří koreferenčních výrazů A, B a C a anotátor přehlédne B a spojí koreferenční šipkou C s A (obr. 44, krok 1.),

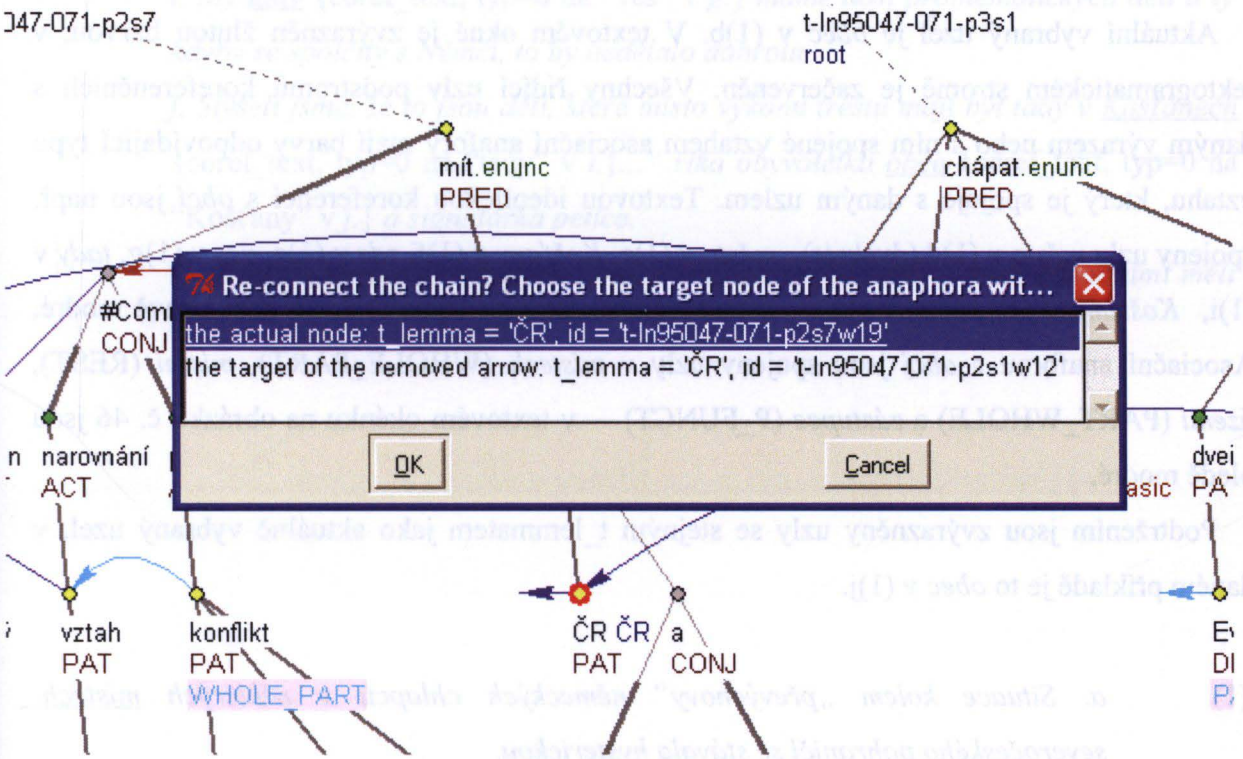
[119] NE = named entity = pojmenovaná entita.

potom se však rozhodne odkázat B na A, automatický nástroj přesměřuje šipku C → A na C → B (obr. 44, krok 2.). Vytvoří se tedy správný koreferenční řetězec A → B → C:



Obrázek č. 44: Vyhledávání nejbližšího antecedentu

3. Dodržování koreferenčního řetězce. Pokud anotátor smaže koreferenční vztah a přeruší přitom koreferenční řetězec, je tázán, chce-li daný koreferenční řetězec skutečně přerušit, nebo pouze „vyndat" ze zachovaného řetězce jeden uzel. Srov. obrázek č. 45 zobrazující situaci, kdy se anotátor pokouší smazat koreferenční šipku vedoucí od *ČR* k *ČR* v předcházejícím kontextu:



Obrázek č. 45: Dodržování koreferenčního řetězce

4. Zdůrazňování výrazů v textu. Anotace rozšířené koreference a asociační anafory probíhá na tektogramatické rovině PTD. Avšak anotátoři mohou pracovat i na povrchové

textové reprezentaci a tektogramatické stromy používat jenom v některých případech. Při výběru výrazu ve větě (kliknutím) nástroj zobrazí, který uzel v tektogramatickém stromě mu odpovídá. Dále na povrchové textové rovině se ukazují již zaznamenané vztahy gramatické a textové koreference a asociační anafory – výrazy, které jsou spojené s jedním z označovaných vztahů mají odpovídající barvu (gramatická koreference – červená, textová koreference – tmavě modrá, asociační anafora – bledě modrá), přičemž asociační anafora se ukazuje i v případě, že je uzel spojen vztahem asociační anafory s jedním z elementů koreferenčního řetězce, jehož je daný uzel součástí. Kromě toho podtržením se zvýrazní všechny uzly, které mají stejné t_lemma jako vybraný uzel.

Na obrázku č. 46 je zobrazen snímek obrazovky anotačního nástroje TrEd v režimu anotace rozšířené textové koreference a asociační anafory (viz v pravé horní části Context PML_T_Bridging a Style PML_T_Bridging). Anotuje se kontext (1)a–k. Horní polovina obrazovky je textové okno, kde je zobrazena aktuální věta (1)b a její předchozí a následující kontext (v daném případě – jedna věta předcházející (1)a a 12 následujících vět (1)c–k). V dolní polovině obrazovky je tektogramatická struktura aktuální věty a jejího nejbližšího kontextu (v daném případě dva stromy před aktuálním stromem a jeden následující strom).[120]

Aktuální vybraný uzel je *obec* v (1)b. V textovém okně je zvýrazněn žlutou barvou, v tektogramatickém stromě je začerveněn. Všechny řídící uzly podstromů koreferenčních s daným výrazem nebo s ním spojené vztahem asociační anafory mají barvy odpovídající typu vztahu, který je spojuje s daným uzlem. Textovou identickou koreferencí s *obcí* jsou např. spojeny uzly *město* v (1)d (dvakrát), *město* v (1)e, *Košťany* v (1)f, *zde* v (1)g, *ves* v (1)g, *tady* v (1)i, *Košťany* v (1)j, *obec* v (1)j – v textovém okénku na obrázku č. 46 jsou tmavě modré. Asociační anaforou s *obcí* jsou spojeny uzly *v místech* (WHOLE_PART), *místní* (REST), *území* (PART_WHOLE) a *zástupce* (P_FUNCT) – v textovém okénku na obrázku č. 46 jsou bledě modré.

Podtržením jsou zvýrazněny uzly se stejným t_lemmatem jako aktuálně vybraný uzel, v daném příkladě je to *obec* v (1)j.

(1) *a. Situace kolem „převýchovy" německých chlapců v některých místech severočeského pohraničí se stávala hysterickou.*
b. V obci {bridging, typ=WHOLE_PART na "místo" v *a.*} *Košťany na Teplicku ještě chlapci ani nebyli, ale místní* {bridging, typ=REST na "obec"} *již dali dohromady petici: „My rodiče dětí základní školy Košťany protestujeme proti*

[120] Všechny parametry v zobrazení stromů a textového okna jsou nastavitelné.

umístění ubytovny pro potrestané německé chlapce.

c. Víme, že na našem území {bridging, typ=PART_WHOLE na "obec" v *b.*} *páchali dál trestnou činnost a nevhodně pokřikovali na kolemjdoucí.*

d. Pokud zástupci {bridging, typ=P_FUNCT na "obec" v *b.*} *města* {coref_text, typ=0 na "obec" v *b.*} *neprovedou opatření proti umístění těchto německých chlapců v našem městě* {coref_text, typ=0 na "město" v *d.*}, *nebudeme posílat svoje děti do školy z důvodu strachu o jejich bezpečnost.*

e. Věříme, že nám všem jde o spokojený život v našem městě {coref_text, typ=0 na "město" v *d.*}... *" a skoro tři stovky podpisů.*

f. V Košťanech {coref_text, typ=0 na "město" v *e.*} *totiž zakoupila dům firma Struktura, která se u nás rozmísťováním německých chlapců zabývá.*

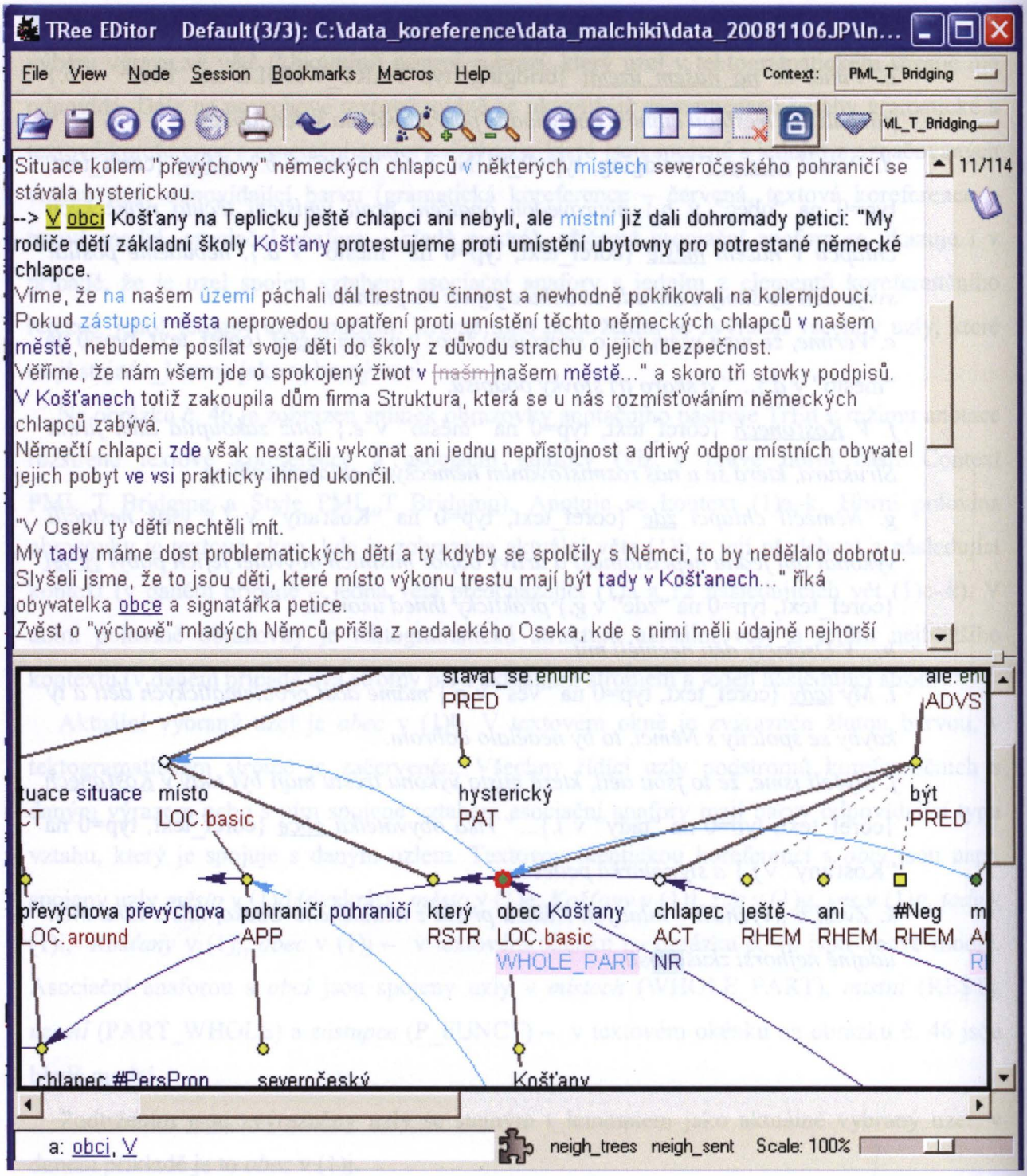
g. Němečtí chlapci zde {coref_text, typ=0 na "Košťany" v *f.*} *však nestačili vykonat ani jednu nepřístojnost a drtivý odpor místních obyvatel jejich pobyt ve vsi* {coref_text, typ=0 na "zde" v *g.*} *prakticky ihned ukončil.*

h. „V Oseku ty děti nechtěli mít.

i. My tady {coref_text, typ=0 na "ves" v *g.*} *máme dost problematických dětí a ty kdyby se spolčily s Němci, to by nedělalo dobrotu.*

j. Slyšeli jsme, že to jsou děti, které místo výkonu trestu mají být tady v Košťanech {coref_text, typ=0 na "tady" v *i.*}... *" říká obyvatelka obce* {coref_text, typ=0 na "Košťany" v *j.*} *a signatářka petice.*

k. Zvěst o „výchově" mladých Němců přišla z nedalekého Oseka, kde s nimi měli údajně nejhorší zkušenosti.



Obrázek č. 46: Zdůrazňování výrazů v textu

IV.2. Aplikace anotace rozšířené textové koreference a asociační anafory

Anotace rozšířené textové koreference a asociační anafory začala v listopadu 2008. Do konce roku 2008 na anotaci pracovali tři anotátoři, lingvisticky vzdělaní studenti FF UK. Od začátku roku 2009 na anotaci pracují dva anotátoři a autorka dané práce.

Anotátoři pracují na různých textech. Každý anotátor anotuje cca. 1000 vět měsíčně, přičemž věty jsou anotovány pouze jednou. Jednou za dva měsíce se provádí test pro měření mezianotátorské shody (viz IV.3.), ve kterém anotátoři dostávají cca. 100 stejných vět.

Statistika oanotovaných dat na 22.12.2009 je představena v tabulce č. 22:

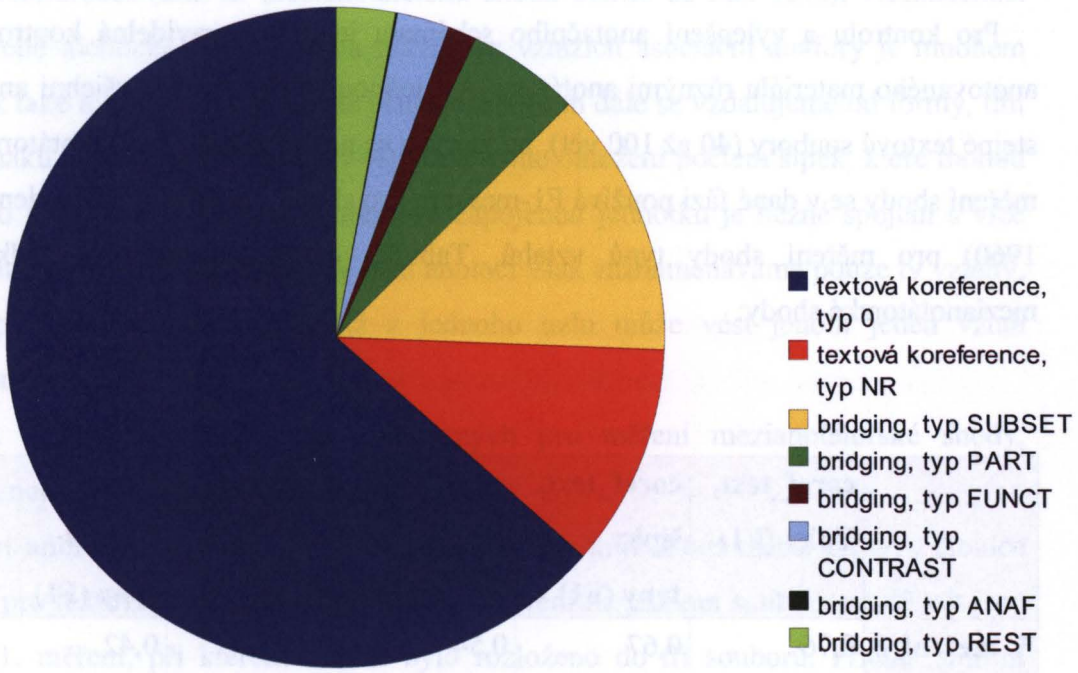
počet anotovaných dokumentů	1580
počet vět	23891
počet slov	403990
počet tektogramatických uzlů (bez technických kořenů stromů)	327380
počet nově anotovaných koreferujících uzlů (rozšířená textová koreference a asociační anafora)	45726
počet původních koreferujících uzlů pronominální textové koreference	10747
počet všech anotovaných koreferujících uzlů (textová koreference (včetně původní pronominální) a asociační anafora)	55744
procento koreferujících uzlů	17,00%
procento anotovaného PDT	50,00%

Tabulka č. 22: Statistické údaje o anotaci textové koreference a asociační anafory na PDT

Tabulka č. 23 a graf č. 4 představuje proporci a statistiku typů textové koreference a asociační anafory anotovaných na PDT.

textová koreference, typ=0 (s pronominální)		16075
textová koreference, typ=0 (bez pronominální)		12034
textová koreference, typ=NR (s pronominální)		2759
textová koreference, typ=NR (bez pronominální)		2740
	textová celkem	**18834**
asociační anafora, typ=SUB_SET		913
asociační anafora, typ=SET_SUB		2394
	celkem SUBSET	3307
asociační anafora, typ=PART_WHOLE		355
asociační anafora, typ=WHOLE_PART		1053
	celkem PART	1408
asociační anafora, typ=P_FUNCT		254
asociační anafora, typ=FUNCT_P		101
	celkem FUNCT	355
asociační anafora, typ=CONTRAST		655
asociační anafora, typ=ANAF		24
asociační anafora, typ=REST		733
	celkem asociační anafora	**6482**

Tabulka č. 23: Statistika typů vztahů textové koreference a asociační anafory



Graf č. 4: Statistika typů textové koreference a asociační anafory

Jak vidíme z diagramu, největší podíl anotovaných vztahů vytváří textová koreference typu 0. Je to podmíněno také skutečností, že do diagramu jsou začleněny vztahy původní pronominální textové koreference, které jsou převážně daného typu. Menší, ale také výraznou skupinu tvoří vztahy textové koreference mezi generickými jmennými frázemi typu NR. V rámci vztahů asociační anafory převážnou většinou tvoří vztahy typu „množina – podmnožina" (3307 vztahů), především z toho důvodu, že při výběru mezi typy SUBSET a PART daný typ je vždy preferován (viz vysvětlení v III.5.1.2.3.1.). Další frekvenční skupinu tvoří typ PART (1408 vztahů). Vztahy typů FUNCT, CONTRAST a REST jsou frekvenčně poměrně vyrovnané. Nejméně je vztahů asociační anafory typu ANAF (24), což se vysvětluje tím, že daná skupina byla zavedena již v průběhu anotace, asi měsíc před tím, než proběhlo dané měření.

IV.3. Měření mezianotátorské shody

Pro kontrolu a vylepšení anotačního schématu je nutná pravidelná kontrola jednotnosti anotovaného materiálu různými anotátory. Asi jednou za dva měsíce všichni anotátoři anotují stejné textové soubory (40 až 100 vět), na kterých se následně měří mezianotátorská shoda. Pro měření shody se v dané fázi používá F1-measure pro shodu v určování antecedentů a κ (Cohen, 1960) pro měření shody typů vztahů. Tabulka č. 24 zobrazuje výsledky tří měření mezianotátorské shody:

	coref_text, šipky (F1)	**coref_text, šipky + typy (F1)**	**coref_text, pouze typy (κ)**	**bridging, šipky (F1)**	**bridging, šipky + typy (F1)**	**bridging, pouze typy (κ)**
1. měření, 3 soubory	0,76	0,67	0,54	0,49	0,42	0,79
2. měření, 40 vět, 1 soubor	0,64	0,41	0,33	0,52	0,52	1
3. měření, 100 vět, 1 soubor	0,80	0,68	0,67	0,59	0,57	0,88

Tabulka č. 24: Výsledky měření mezianotátorské shody

Zvlášť se počítá a) shoda v zaznamenání koreferenční/bridging šipky (1. a 4. sloupec resp.), b) shoda v zaznamenání koreferenční/bridging šipky, a zároveň stejné určení typu vztahu (2. a 5. sloupec resp.) a c) shoda pouze typů vztahu na šipkách označených stejně oběma anotátory (3. a 6. sloupec resp.).

IV.4. K rozdílům v mezianotátorské shodě

Výsledky měření mezianotátorské shody jsou značně různorodé a potřebují další vysvětlení. Z uvedených dat je vidět, že za tři změřená období se mezianotátorská shoda výrazně nezlepšila, dokonce v několika místech se v následujících měřeních zhoršuje.

Úroveň anotace rozšířené koreference a asociační anafory je v podstatě již textová, je to určitá nadstavba nad tektogramatickou rovinou. Anotace v mnoha případech závisí na

interpretaci textu, které může být u různých anotátorů odlišné. Ve srovnání s původní pronominální koreferencí (kde se mezianotátorská shoda měřila až nad 95%), víceznačnost vztahů v rozšířené identické koreferenci a hlavně ve vztazích asociační anafory je mnohem větší. Je to však také logická a odůvodněná skutečnost – čím dále se vzdalujeme od formy, tím je polysémie znaků větší (srov. Melčuk 1974). Také jsme omezeni počtem šipek, které mohou vést od jednoho uzlu. Pro skutečnou kontextově zapojenou jednotku je běžné spojení s více jinými jednotkami různými typy vztahů. V naší anotaci však zaznamenáváme pouze ty vztahy, které jsme explicitně definovali, přičemž z jednoho uzlu může vést jenom jeden vztah asociační anafory (viz III.1.8.).

Jak ukázala detailnější analýza dat anotovaných pro měření mezianotátorské shody, výsledek shody nejvíce ovlivňují následující faktory:

1. **Velikost anotovaného souboru**. Čím delší je text, tím je shoda menší. Srov. v tabulce č. 24 výsledky pro textovou koreferenci při druhém měření na jednom souboru ze 40 vět jsou horší než pro 1. měření, při kterém 40 vět bylo rozloženo do tří souborů. Příčina snížení mezianotátorské shody je především v tom, že u delších textů je síť koherenčních vztahů delší, a tedy výrazně složitější a mnohoznačnější; při vyhledávání antecedentů má tedy anotátor více možností. Na druhé straně komplikovaná koherenční struktura vede k tomu, že některé vztahy uniknou pozornosti anotátora (viz rozbor textu dále v této kapitole).
2. **Úroveň abstrakce anotovaného textu**. Čím více je v textu jmen s abstraktním významem a výrazů s generickou referencí, tím je shoda menší.

Obě kritéria mají na mezianotátorskou shodu dramatický vliv. Je-li text stručný a konkrétní, mezianotátrská shoda je téměř stoprocentní. Jakmile se do textu zapojí abstraktnější pojmy a obecnější tvrzení, shoda prudce klesá.

V následujícím výkladu jsme podrobili detailní klasifikační analýze výsledky mezianotátorské shody na jednom větším souboru (101 vět). Náš rozbor má však pouze exemplifikační platnost – výsledky této analýzy nebyly srovnány s anotacemi jiných souborů, ani jsme nevytvářeli statistiku chyb. Vybrali jsme pouze pro ukázku ty nejtypičtější případy neshody, se kterými se setkáváme vždy, když srovnáváme různé anotace jednoho souboru.

Tři anotátoři anotovali následující text (2):

(2) *(1) Děti, rodina a stres.*
(2) Vladislav Vít.

(3) Před prázdninami vyšla v nakladatelství Galén kniha Děti, rodina a stres s podtitulem Vybrané kapitoly z prevence psychické zátěže u dětí.

(4) Její autoři prof. PhDr. Zdeněk Matějček a Doc. MUDr. Zdeněk Dytrych před pětadvaceti lety založili Oddělení pro výzkum rodiny, které dodnes vedou.

(5) Z. Matějček se věnuje dětem a Z. Dytrych dospělé části rodiny.

(6) Zdeněk Dytrych: Na ministerstvu zdravotnictví v útvaru hlavního hygienika se objevila potřeba shrnout některé problémy, které se v rodině velice často opakují.

(7) Profesor Matějček a já jsme byli požádáni, abychom se o takové shrnutí pokusili s tím, že čtenáři mají dostat konkrétní rady.

(8) Tak je i knížka koncipována.

(9) V každé kapitole se mluví o určitém problému, uvádíme jak je rozsáhlý, kolik dětí je jím postiženo a co dělat.

(10) Je tam v podstatě konkrétní návod.

(11) Vaše kniha obsahuje ve třiadvaceti kapitolách různé problémy, od těžkých poškození dítěte až po lehčí disfunkci či vliv rozvodu na dítě.

(12) Tím ovšem jednu konkrétní rodinu může zajímat maximálně pět, přinejhorším deset kapitol.

(13) Zdeněk Matějček: Původně tato knížka byla určena pro zdravotnické pracovníky, a to především pro lékaře, kteří jsou ve styku s rodinou.

(14) Na druhé straně se ukázalo, že toto téma je stejně důležité pro pedagogy a vychovatele.

(15) Ti se přece setkávají i s postiženými nebo týranými dětmi.

(16) A když už byla knížka hotova, tak se zjistilo, že je praktická i pro rodiče.

(17) Samozřejmě ne každá kapitola ne pro každého rodiče.

(18) Zdeněk Dytrych: Kdyby se přímo dotýkalo některé rodiny deset kapitol, tak by to byla opravdu nešťastná rodina.

(19) Ale stačí jedna a většinou jich bude i víc.

(20) Vezměte si, kolik je rozvodů – třicet tisíc ročně v republice, to znamená, téměř třicet tisíc dětí je rozvodem nějakým způsobem postiženo.

(21) V této knize je poučení, jak snášejí děti rozvod a jak na něj reagují, a návod, jak se mají rodiče chovat, aby se utrpení dětí snížilo.

(22) Nebo například existuje lehká mozková disfunkce, kterou trpí podle našeho rozsáhlého výzkumu pět procent dětí.

(23) Toto postižení se velice špatně rozpoznává.

(24) Dítě je nemotorné, neklidné a není schopné se soustředit, ale přitom je většinou chytré.

(25) Rodiče ho považují za lajdáka a bývá trestáno třeba za špatný výkon ve škole, tím se zhoršuje vztah k učení atd.

(26) A tohle rodiče musí vědět.

(27) Samozřejmě i pedagogové a v této knížce je návod co s tím.

(28) Zdeněk Matějček: Předkládáme i problémy, na které se zapomíná.

(29) Tak například úmrtí dítěte nebo narození postiženého dítěte.

(30) Tady nejde jenom o rodiče, ale i o okolí, které musí vědět, jak se má chovat.

(31) Nebo úmrtí v rodině a jeho vliv na dítě a může to být třeba babička.

(32) Dá se říci, že kapitoly z vaší knížky, které se určité rodiny netýkají, přispějí k porozumění těm druhým?

(33) Zdeněk Matějček: Ano, to je přesné.

(34) Vždyť například chlapec s lehkou mozkovou disfunkcí není jen v rodině či ve škole.

(35) Taková rodina má své přátele, sousedy atd.

(36) My ovšem touto knížkou nemůžeme dát přesný recept, ale to, co zdůrazňujeme, je porozumění čili rozumět tomu a pak se prostředek k nápravě najde.

(37) A když porozumím předem, tak i mohu předcházet.

(38) Materiálům, které dnes máte k dispozici, předcházel dlouholetý výzkum.

(39) Zdeněk Dytrych: Od roku 1969, kdy jsme založili v bývalém Výzkumném ústavu psychiatrickém Oddělení pro výzkum rodiny, se hlavně zabýváme touto problematikou.

(40) Měli jsme samozřejmě řadu spolupracovníků a za pětadvacet let jsme v týmu udělali téměř nekonečnou řadu prací.

(41) Tak například rozsáhlý výzkum rozvodovosti.

(42) Sledovali jsme šest let po rozvodu, co se děje s bývalými manžely a s jejich dětmi.

(43) Jaké je jejich další začlenění do života, jak je rozvod poznamenal, či nepoznamenal.

(44) Zdeněk Matějček: Studovali jsme děti, které vyrůstaly v situacích za méně příznivých společenských a citových podmínek, které mohou být nejrůznějšího druhu.

(45) Kdy hned na začátku do osudu dítěte vstoupilo něco nedobrého.

(46) Jsou to děti, které získaly určité obtíže z vnějšího světa.

(47) Na druhé straně se zabýváme dětmi, které si nesou v sobě od počátku určitou nápadnost nebo obtíž.

(48) Zdeněk Dytrych: K tomu je třeba dodat, že za poslední léta v Čechách roste počet dětí s vrozenou vývojovou vadou.

(49) My jsme tento nárůst sledovali a je to až do deseti procent.

(50) Vybrali jsme si jako modelovou situaci rozštěp rtu a patra.

(51) Zkoumali jsme, co se děje s matkou.

(52) Ta zažije šok a táhne se to s ní celý život a s dítětem taky.

(53) Jsou případy, kdy se matka zavírala doma a přerušila styky s celým okolím i s přáteli.

(54) To dítě je totiž tzv. nechlubitelné.

(55) Rodina se stáhne do sebe a vzniká zvláštní atmosféra, která samozřejmě vytváří stres.

(56) Má to různé fáze a ty my sledujeme, podobně jako lékaři ve Spojených státech a jinde.

(57) Zdeněk Matějček:

(58) Náš výzkum bych shrnul do dvou kategorií.

(59) Děti, které si nesou do života problém, který není nápadný, který je uložený v jeho centrálním nervovém systému, jak už vrozeně, tak vlivem vnějším.

(60) A děti, které mají nápadnou vadu a můžete je sledovat takřikajíc jako na dlani.

(61) Zdeněk Dytrych: Některé výzkumy, ty největší, trvají pětadvacet let a vyžadují obrovské úsilí.

(62) Zdeněk Matějček: Jeden projekt, který ještě běží, je sledovat od určitého věku děti, které jsou dnes dospělé a mají děti, ale vyrostly za poněkud zvláštních podmínek.

(63) Jedna skupina jsou jedinci, kteří nikdy nepoznali rodinu a vyrůstali v dětských domovech.

(64) V další skupině jsou ti, kteří z dětských domovů přešli do náhradní péče, ale typu SOS vesničky, a v třetí ti, kteří se dostali do individuální pěstounské péče.

(65) Dále je skupina, která vyrůstala v rodině vlastní za méně příznivých podmínek (např. nechtěné děti) a nakonec je skupina kontrolní, která reprezentuje tzv.

normál.

(66) Jsou nějaké hlavní cíle?

(67) Zdeněk Dytrych: Jsou to dvě věci.

(68) Jedna je tzv. patologie třetí generace.

(69) Například: jestliže matka nechtěla dítě a dítě se jí narodilo proti její vůli, vyvíjelo se nepříznivě a je vysoká pravděpodobnost, a my chceme vědět jaká, že i toto dítě se v budoucnu bude chovat ke svému dítěti podobně.

(70) Koneckonců tento vysoce zajímavý jev byl pozorován v experimentech u opic, které prováděl před několika desítkami let vynikající americký etolog Harlow.

(71) Opice, kterým se nedostalo mateřské lásky, poněvadž po narození jejich matka byla nahrazena kovovou atrapou, ve své dospělosti pak zavrhovaly svoje mláďata.

(72) Zdeněk Matějček: Jak dokazuje výzkum z dětských domovů, když jsou tyto děti dospělé, tak nejsou schopny lásku přijímat ani dávat.

(73) Zdeněk Dytrych: Čili do koho nebyla láska vložena, tak ji nedovede ani dávat.

(74) Zdeněk Matějček: Tento člověk není schopen citové investice a je s ním život těžší;

(75) ne že by to bylo prokletí a že by byli všichni takoví, ale je to tendence.

(76) Prokáže se, když sledujete větší skupiny.

(77) Zdeněk Dytrych: Chceme znát možnosti, jak tomu zabránit;

(78) a to souvisí i s druhou částí, kdy hledáme faktory tzv. protektivní, které vedou k tomu dobrému ve vývoji.

(79) Zdeněk Matějček: Přesto jsou děti, které v takovýchto podmínkách vyrostou v dobré, bezproblémové dospělé.

(80) Proč?

(81) Nejde jen o to hledat negativa, ale i pozitiva.

(82) Co to způsobilo?

(83) Například to může být škola, sport, hudba atd.

(84) Většinou něco, co jim umožnilo vyniknout.

(85) Zdeněk Dytrych: Tyto děti, které prožily některé období naplněné stresem, mají společný znak a to je ztráta sebedůvěry.

(86) Mají ji rozkolísanou.

(87) Nevěří v sebe, nevěří v ostatní a jejich život se klikatí, má různé složitosti a ty se na sebe hromadí.

(88) My tomu říkáme efekt sněhové koule.

(89) Na negativní věci se nabalují další a další a už ani nevíme, co bylo na začátku.

(90) Mohl to být třeba nedostatek lásky rodičů, především matky.

(91) Zdeněk Matějček: Když takové dítě někde vynikne, tak se vlastně efekt sněhové koule obrátí.

(92) Začne se rozvinovat.

(93) Setkáváte se s dětmi, které mají ve škole problémy se svou přebujelou fantazií?

(94) Zdeněk Matějček: Není to časté, ale chodí k nám rodiče takovýchto dětí a žádají vyšetření.

(95) Přicházejí děti, kde je přebujelá fantazie jen mírně naznačená, ty, kde je výrazná, až po ty, které nemohou z tohoto důvodu chodit do školy.

(96) Musí se učit doma.

(97) Zrovna nyní mám jeden případ, kde se pokoušíme zkusit to ve čtvrté třídě.

(98) Aby aspoň nějakou hodinu ve škole byl.

(99) Zdeněk Dytrych: Není to přesně, na co se ptáte, ale problém s fantazií není jen, že ji má dítě vrozenou, ale jde o únik z reality, která je pro dítě málo přijatelná.

(100) Vysnívá si jiný svět.

(101) A to je většinou zase problém rodiny.

Při rozboru tří variant anotace vyšly najevo následující typy neshod mezi anotátory:

1. **Různá chápání textu**. Srov. např. vztahy mezi výrazy v (3):

(3) *knížka (8) – kapitola (9) – tam (10)*

Výraz *tam* v 2.10 anotátor A odkázal jako coref_text, typ=0 na *kapitola* v 2.9, zatímco anotátor B – jako coref_text, typ=0 na *knížka* v 2.8. Obě interpretace jsou stejně možné a vedou ke stejné interpretaci textu jako celku. Je to případ více technických možností odkazování, které de facto nejsou spojeny s porozuměním daného textu. S velkým překvapením jsme zjistili, že podobných příkladů se při rozboru anotovaných koreferenčních vztahů nachází

výrazně více, než jsme očekávali. Takové případy však není možné odstranit zdokonalením anotačního schématu ani automatickou předanotací, naše mezianotátorská shoda tedy nebude nikdy stoprocentní.

Příklady různého chápání textu jsou následující:

(4) *vztah dlouholetý výzkum (38) – řada prací (40) – rozsáhlý výzkum rozvodovosti (41)*

V (4) anotátor A vedl šipku vztahu bridging_SET_SUB na *dlouholetý výzkum*, anotátor B – na *práce*. Obě interpretace jsou možné a stejně pravděpodobné: rozsáhlý výzkum rozvodovosti je na jedné straně součástí dlouholetého výzkumu, o kterém se hovoří v celém článku a který je zmíněn v 2.38, na druhé straně je to také jednotlivý příklad nekonečné řady prací, které jsou rématem v právě předcházející větě 2.40.

V příkladu (5) je ukázána neshoda ve výběru antecedentu.

(5) *patologie třetí generace (68) – jestliže matka nechtěla dítě a dítě se jí narodilo proti její vůli, vyvíjelo se nepříznivě a je vysoká pravděpodobnost, a my chceme vědět jaká, že i toto dítě se v budoucnu bude chovat ke svému dítěti podobně (69) – tento vysoce zajímavý jev (70)*

Anotátor A za antecedent NP *tento vysoce zajímavý jev* považuje generické pojmenování celého jevu – *patologie třetí generace* a odkazuje na něj textovou generickou koreferencí (typ=NR). Podle anotátora B antecedentem NP *tento vysoce zajímavý jev* je konkrétní příklad této patologie popsaný celou větou 2.69. Obě varianty koreferenčního vztahu vedou ke stejné interpretaci referenční platnosti anaforického výrazu a textu jako celku.

Srov. také (6):

(6) *negativa (81) – složitosti (87) – negativní věci (89)*

Jeden z anotátorů označil vztah mezi *negativa* a *negativní věci* jako textovou koreferenci typu NR. Není to vyloučeno, je to dokonce poměrně logické. Je možné jej pochopit různě, a dokonce nemusí být ani za koreferenci považován. Druhý anotátor odkázal *negativní věci* na *složitosti* v 2.87, což také není vyloučeno, ale podobně jako první řešení je i toto nejednoznačné.

2. **Rozhodování mezi odkazem na segment textu** (coref_special, typ=segm) **a vztahem s explicitním antecedentem** (coref_text nebo bridging). Poměrně často jeden anotátor vidí odkaz na větší úsek textu (a anotuje ho jako coref_special, typ=segm) tam, kde jiný anotátor je schopen najít explicitní antecedent.

V příkladě (7) anaforické hyperonymum *toto téma* v 2.14 mělo tři různé anaforické interpretace. Anotátor A odkázal *toto téma* na *tato knížka* v 2.13 vztahem bridging_REST (jako *kniha – téma*), anotátor B spojil textovou koreferencí *toto téma* a *problémy* v 2.11, anotátor C odkázal *toto téma* na větší předcházející kontext (coref_segm). Všechny tři vztahy jsou diskutabilní a všechny jsou teoreticky možné.

(7) *problémy, od těžkých poškození dítěte až po lehčí disfunkci či vliv rozvodu na dítě (11) – tato knížka (13) – toto téma (14)*

Srov. také (8), kde se možný antecedent textové koreference nachází ve vzdálenosti dvacet vět od anaforu. Anotátor A v páru *dlouholetý výzkum – náš výzkum* označil textovou koreferencí typu 0, anotátor B odkázal *náš výzkum* v 2.58. speciální koreferencí typu segm na celý předcházející kontext. *Výzkum* v analyzovaném textu má hypertematickou platnost, tedy i v 2.58 je kontextově zapojený. Tato kontextová zapojenost se však zachovává jinými vztahy, než textovou koreferencí výrazů v (8).

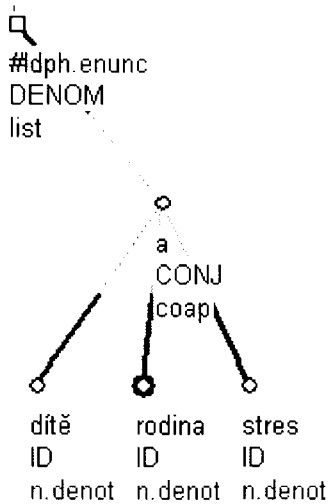
(8) *dlouholetý výzkum (38) – náš výzkum (58)*

S velkou pravděpodobností tu jde o skutečnou koreferenci mezi *dlouholetý výzkum* v 2.38 a *náš výzkum* v (8) tedy logicky je správné řešení anotátora A, ale velká vzdálenost antecedentu v textu a zaměření celého předcházejícího kontextu 2.39–2.57 na výzkum vysvětluje i druhé řešení.

3. **Technické nesrovnalosti – konvence proti intuici.**

a) **Hloubka anotovaného uzlu.** V tomto případě dochází k rozporu konvence (princip maximální velikosti koreferovaných členů) a intuice anotátora. V příkladě (9) se konvenční řešení (odkazovat na uzel s t-lemmatem #Idph) zdá být kontraintuitivní (srov. obrázek č. 47), proto jeden z anotátorů odkázal NP *kniha Děti, rodina a stres* v 2.3) na významový kořen, čili na spojku „a“, nikoliv na #Idph:

(9) *Děti, rodina a stres... (1) – kniha Děti, rodina a stres (3)*



Obrázek č. 47: Antecedentní věta 2.1

b) **Anotace uzlů s funktorem ID nebo částí pojmenovaných entit.** Konvence neanotace uzlů s funktorem ID (III.4.2.2.3.) a částí pojmenovaných entit, které samy nejsou pojmenovanými entitami (III.4.2.2.3.2.) nejsou vždy anotátory dodrženy, pokud se dále v textu vyskytnou anaforické odkazy na tyto entity. Srov. v (10) jeden z anotátorů označil koreferenční vztah mezi *rodina* v 2.4 a *rodina* v 2.5, i když v *rodina* v 2.3 má funktor ID a *rodina* v 2.4 je součástí pojmenované entity.

(10) *kniha Děti* . ID, *rodina.ID a stres.* ID *(3) – Oddělení pro výzkum rodiny (4)*

V daném případě existující konvence částečně protiřečí intuici anotátora a rovněž skutečnosti, že vztah mezi NP *rodina* v 2.4. a 2.3. je (genericky) koreferenční a podílí se na kohezi textu.

4. Hranice mezi asociační anaforou (hlavně typu SUBSET) a koreferencí typu NR. Jde především o vztah mezi nereferenčně (genericky) použitými, často hypertematickými jmennými frázemi. V takových případech, zvlášť u dlouhých koherenčních posloupností, je někdy obtížně rozhodnout mezi textovou koreferencí typu NR a vztahem asociační anafory typu SUBSET.

Srov. např. vztah mezi dvěma genericky použitými NP *rodiče* v 2.16 a 2.21. V 2.16 se však mluví o všech rodičích, zatímco v 2.21 – pouze o rodičích postižených rozvodem, což je podmnožina množiny všech rodičů v 2.21. Generická referenční platnost obou výrazů v 2.16 a 2.21 však ukazuje také ke koreferenční interpretaci typu NR. Anotátor A vybral asociační anaforu typu SUBSET, anotátor B – textovou koreferenci typu NR.

Srov. také podobné vztahy mezi *matka* v 2.51 a v 2.53 a *dítě* v 2.52 a 2.54. V obou větách 2.51 a 2.53 je *matka* použita nereferenčně, avšak stále se jedná o matky dětí s rozštěpem rtu a patra. V podstatě se tvrdí, že některé matky takto nemocných dětí se zavírají doma atd. Logicky správný je tedy vztah asociační anafory typu SUB_SET, ale generická reference opakujících se stejných NP odkazuje také k textové koreferenci typu NR. Anotátor A provedl textově koreferenční vztah typu NR, anotátor B – asociační anaforu typu SET_SUB. Velice podobně vypadají referenční vztahy v páru *dítě – dítě* ve větách v 2.52 a 2.54. Nechlubitelné je kterékoli postižené dítě, nejenom u té matky, která se zavřela doma a přerušila styky s celým okolím i s přáteli (věta 2.53). I když takové chápání rovněž není vyloučeno. Obě řešení jsou logicky oprávněná. Konvence preference koreference před asociační anaforou vede v tomto případě k označení koreference typu NR. Paralelismus konstrukcí 2.51 + 2.53 a dítě v 2.52 + 2.54 a jejich sémantiky je zřejmý. Také je zřejmé, že odlišná interpretace koreferenčních vztahů v takových případech je velice časově náročná a není zaručeně použitelná.

Celkově je poměrně nepředvídatelná shoda u vztahů asociační anafory mezi jmennými frázemi s nespecifickou referencí – srov. velký počet možností navazování koreferenčních vztahů mezi jmennými frázemi s podtrženými řídícími uzly *dítě, rodič* a *rodina* ve větách 2.5–2.35.

Celý rozhovor (2) probíhá na abstraktní rovině, tj. nemluví se o konkrétních dětech ani rodinách. Není pravda, že všechny NP *rodiče* a všechny NP *děti* jsou koreferenční. V některých větách je zřejmý vztah SUBSET (*Matějček se věnuje dětem – kolik dětí je jím postiženo; Dytrych dospělé části rodiny – jednu konkrétní rodinu může zajímat...* atd). Avšak jasné příklady jsou bohužel vždycky v menšině. Např. dále ve větě 2.11 už je pro vztah mezi dvěma výrazy *dítě* více možností: koreferovat *dítě* na předcházející *dítě* v obecněgenerickém kontextu, nebo je spojit asociační anaforou typu SUBSET. Řešení podobných problémů je neproduktivní ztrátou času a jen stěží k něčemu přispěje. Proto je v daném případě pravděpodobně lepším řešením koreferovat podobné uzly s typem NR. Problém je však v tom, že jen velice těžko se stanoví omezení takového rozhodnutí.

Argumenty pro zaznamenání textové koreference typu NR a asociační anafory typu SUBSET a jejich nevýhody jsou shrnuty v tabulce č. 25:

řešení	**opodstatnění**	**nevýhody**
coref_text, typ=NR	• propojíme všechny související generické výrazy do jednoho řetězce, protože skutečně přispívají ke kohezi textu • je to jednodušší a časově méně náročný přístup • je myslitelná případná automatizace	• kontraintuintivní chápání koreference • mohou se ztratit skutečné koreferenční řetězce • rozhodneme-li se pro coref_text, typ=NR, budeme se muset držet tohoto přístupu i v případě jasných vztahů SUBSET v neambiguitních kontextech generických NP
asociační anafora, typ=SUBSET	• jmenné fráze ve skutečnosti nereferují na stejnou množinu objektů, typ=SUBSET více odpovídá reálné situaci • v jednodušších kontextech, kde je vztah SUBSET zřejmý, není třeba zvažovat jeho označení	• složitost až nemožnost takovým způsobem oanotovaná data jakýmkoliv způsobem automaticky zpracovat • časová náročnost • nízká mezianotátorská shoda

Tabulka č. 25: Hranice mezi asociační anaforou (hlavně typu SUBSET) a textovou koreferencí typu NR

V dané fázi anotace se přikláníme k anotaci podobné ambiguity podle smyslu, přičemž v případě váhání anotátor vybere textovou koreferenci typu NR, ale je obtížné stanovit exaktní pravidla.

5. Hranice mezi asociační anaforou SUBSET a nezaznamenáním vztahu. Další specifikací neshody popsané v předchozím bodě je výběr mezi neoznačením žádného vztahu a

označením asociační anafory typu SUBSET. Srov. vztah mezi *dítě* v 2.31 a *děti* v 2.42 a 2.44 po značném, ale menším než hraničních 20 vět, textovém intervalu. NP *děti* v 2.44 nemohou být koreferovány s „dětmi" v 2.42– 2.43, protože jde o jinou problémovou skupinu. Avšak *děti* v 2.44 je jistá podmnožina všech dětí, a všechny děti se už genericky v předchozím textu vyskytly v 2.31, i když ne zcela jednoznačně. Z toho důvodu anotátor A vede asociační anafora typu SET_SUB na „dětí" v 2.31, zatímco anotátoři B a C žádnou šipku nevedou. Podobný problém může vzniknout i při koreferenci NP *děti* v 2.42, kde *děti* mohou být také chápány jako podmnožina všech dětí a tudíž odkázána na *děti* v 2.31.

6. **Hranice mezi textovou koreferencí typu NR a nezaznamenáním vztahu**. Podobně jako ve dvou předcházejících bodech jde o případy, kdy extenze použitých generických substantiv není totožná. Srov. možnost spojit/nespojit generickou koreferencí NP *dětský domov* ve větách 2.63 a 2.64. *Dětský domov* v 2.63 a 2.64 má nejednoznačnou referenci, která může být interpretována i jako specifická. V tom případě by dané NP neměly být spojené koreferenční šipkou. V případě nereferenční interpretace je koreferenční vztah žádoucí.

7. **Asociační anafora typu SUBSET u nereferenčních NP.** Jakmile generické jmenné fráze mají různou extenci, k problematice rozhodnutí označovat/neoznačovat se připojuje také problém správného (resp. stejného u více anotátorů) výběru antecedentu. Srov. v kontextu 2.72–2.79 se hovoří o dětech v dětských domovech, které v dospělosti „nejsou schopny lásku přijímat ani dávat". O takových dětech text pojednával již nějakou dobu. Pak následuje věta (2.79) *„Přesto jsou děti, které v takovýchto podmínkách vyrostou v dobré, bezproblémové dospělé"* V předchozím kontextu se nevyskytla NP s generickým významem, odpovídající denotaci 'všechny děti, které vyrůstají v dětských domovech', tedy nemůžeme z této poslední věty vést asociační anaforu typu SUBSET. Jednoznačně však jsou tyto děti podmnožinou dětí z dětských domovů. Výsledek anotace tedy je, že tuto NP povede každý anotátor někam jinam. Anotátor A vede šipku asociační anafory SUBSET na vzdálenější předcházející kontext, kde se vyskytla NP *děti* v generičtějším významu *(matka nechtěla dítě* v 2.69*)*, anotátor B – bridging_CONTRAST na *tento člověk* v 2.74, anotátor C – bridging SUBSET na *tyto děti* v 2.72.

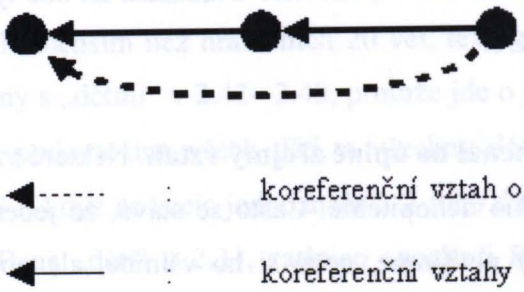
Pokračování daného textového úseku předkládá další problém. Srov. bezprostředně následující kontext 2.80–2.85, kde NP *tyto děti* v 2.85 odkazuje genericky na děti, které vyrostly v dětských domovech a v předchozím kontextu nemá antecedent. Anotátoři ji odkázali také různě: anotátor A odkázal textovou koreferencí typu NR na *děti* v 2.72, anotátor B –

textovou koreferencí typu NR na *šťastné děti* v 2.79, anotátor C odkázal na obě tyto NP (*děti* v 2.72 a v 2.79) vztahem bridging_SUB_SET.

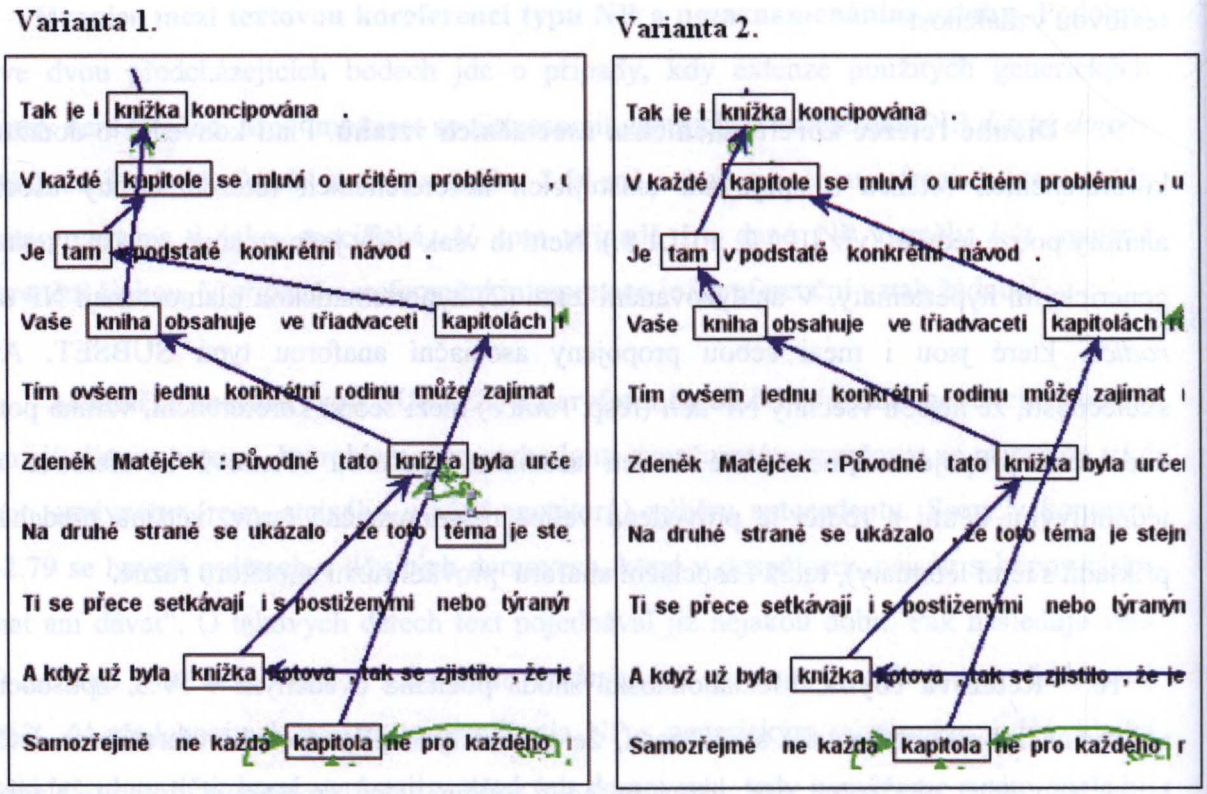
8. Anotátor **zaznamenal/nezaznamenal ne úplně zřejmý vztah.** Některé vztahy (zvláště z oblasti asociační anafory) nejsou snadno uchopitelné. Často se stává, že jeden z anotátorů anotuje vztah, kterého si druhý anotátor nevšimne, nebo si ho všimne, ale považuje ho za nerelevantní. Srov. např. vztah mezi *konkrétní rady* a *konkrétní návod* ve větách 2.7 a 2.10, kde jeden z anotátorů označil vztah mezi *rada* a *návod* jako asociační anaforu typu SET_SUB, zatímco druhý anotátor si ho vůbec nevšiml. Zvlášť aktuální je podobný problém na větší textovou vzdálenost.

9. Dlouhé řetězce koreferenčních a asociačních vztahů. Platí konvence o dodržování koreferenčních řetězců a spojování existujících koreferenčních řetězců vztahy asociační anafory pouze jednou (viz III.1.2. a III.1.8.). Není to však vždy jednoznačné, zvlášť v textech s generickými hypertématy. V analyzovaném textu (2) hypertematickou platnost mají NP *děti* a *rodiče,* které jsou i mezi sebou propojeny asociační anaforou typu SUBSET. Avšak skutečností, že nejsou všechny NP *děti* (resp. *rodiče*) mezi sebou koreferenční, vzniká potřeba dodatečně propojovat jiné rodiče a děti asociační anaforou. Textová koreference mezi jednotlivými dětmi a rodiči je provedena velice nejednoznačně (srov. většinu předchozích příkladů s těmi lemmaty), tudíž i asociační anaforu provádí různí anotátořo různě.

10. Řetězová chyba. Mezianotátoská shoda počítána uvedeným v IV.3. způsobem je snižována kromě jiného také skutečností, že v některých případech koreferenčních řetězců, které jsou delší než jeden pár, jedna neshoda automaticky zapříčiňuje druhou. Například neshoda (3) popsaná v bodu 1. zapříčiňuje další neshodu, protože *kapitola* v 2.11 jednoznačně koreferuje s *kapitola* v 2.9. pokud však na ni už vede *tam* v 2.10, šipka z 2.11 automaticky povede na *tam* v 2.10, nikoliv na *kapitola* v 2.9, jak to bude v případě *tam* v 2.10 s *knížka* v 2.8. Paralelně vzniká druhá neshoda, kde *kniha* v 2.11 jednoznačně koreferenční s *knížkou* v 2.8 bude v jednom případě spojena s *tam* v 2.10, jednou – s *knížkou* v 2.8. Srov. záznam levých částí vět 2.8–2.17 na obrázku č. 48 a schematickou představu typu chyby na obrázku č. 49.



Obrázek č. 48: Řetězová chyba (jedna neshoda vleče druhou)



Obrázek č. 49: Řetězová chyba (jedna neshoda vleče druhou)

V. Závěrem

V této práci jsme představili jeden z možných modelů zpracovaní rozšířené textové koreference a asociační anafory na velkém korpusu textů, který dále používáme pro anotaci daných vztahů na textech Pražského závislostního korpusu.

Na základě literatury z oblastí teorie reference, diskurzu a některých poznatků teoretické lingvistiky na jedné straně a s použitím existujících anotačních metodik na straně druhé, jsme se pokusili vytvořit poměrně detailní klasifikaci textově koreferenčních vztahů a vztahů asociační anafory. Velkou roli při vytvoření klasifikace anaforických a koreferenčních vztahů sehrál rovněž samotný jazykový materiál – četné pokusy s různými hypotézami na skutečných textech rychleji odhalují chyby v klasifikaci, což považujeme za zřejmou výhodu našeho přístupu vůči čistě teoretickým přístupům.

Jedním z úkolů výzkumu bylo vytvořit systém teoretických principů, které je nutno dodržovat při anotaci koreferenčních vztahů a asociační anafory. Jsou to například princip důslednosti anotace, princip dodržování maximálního koreferenčního řetězce, princip kooperace se syntaktickou strukturou tektogramatické roviny, princip preference koreferenčního vztahu před asociační anaforou a další. Tyto principy se pak aplikují při formulaci pravidel anotace v typických problematických situacích, ale také je používají anotátoři v konkrétních případech víceznačné nebo problematické interpretace anotovaných vztahů v textu.

Frekvence koreferenčních vztahů a vztahů asociační anafory se liší podle slovnědruhové hodnoty řídícího výrazu účastníka koreferenčního vztahu. Formální charakteristika koreferovaných uzlů ukázala, že nejčastěji do koreferenčních vztahů vstupují jmenné fráze se substantivní hlavou (66,5%) a osobní zájmena (25%). Méně časté, ale také poměrně frekventované jsou členy anaforického vztahu – souřadné nebo seznamové struktury (1,5%). Podle stanovených pravidel do anaforických vztahů mohou vstupovat některá adjektiva a číslovky. Minimálně se označují koreferenční vztahy u sloves.

Při vypracování typologické klasifikace textové koreference a asociační anafory jsme se snažili zachycovat co nejvíce sémantických rozdílů označovaných vztahů, které mohou být užitečné pro případný budoucí lingvistický výzkum anotovaného materiálu. Zároveň jsme dbali na dodržování maximální přesnosti kritérií vymezování jednotlivých typů vztahů, které je nezbytně důležité pro následné automatické zpracování koreference a asociační anafory. Výsledkem takového kompromisu bylo vytvoření dvou typů textově koreferenčních vztahů – koreferenční vztah mezi jmennými frázemi se specifickou referencí (coref_text, typ=0) a

koreferenční vztah mezi jmennými frázemi s nespecifickou, především generickou, referencí (coref_text, typ=NR). Rozdíl koreferenčních vztahů mezi uvedenými skupinami je teoretický zásadní, otvírá však řadu dalších dílčích a teoretických problémů, jako např. problém neostré hranice mezi specifickými a generickými jmennými frázemi, různé stupně generičnosti generických jmenných frází, problém existence koreferenčního vztahu mezi generickými jmennými frázemi různých úrovní abstrakcí a mnoho dalších.

Pro asociační anaforu jsme stanovili šest typů vztahů: vztah PART mezi částí a celkem, vztah SUBSET mezi množinou a podmnožinou/prvkem množiny, vztah FUNCT mezi entitou a unikátní funkcí na této entitě, vztah CONTRAST sémantického a kontextového protikladu, vztah ANAF anaforického odkazování mezi nekoreferenčními entitami a vztah REST pro jiné případy asociační anafory. Rozdíl mezi skupinami SUBSET a PART se ukázal ve velkém počtu kontextů poměrně vágní, proto byl stanoven poněkud arbitrárně. Vztah FUNCT, ačkoliv v anotačních projektech podobných našemu tradičně vystupuje, v našich textech byl málo frekventovaný (5,5% všech vztahů asociační anafory). Vztah sémantického a kontextového protikladu (CONTRAST) se poněkud liší od jiných typů asociační anafory tím, že představuje spíše diskurzní vztah mezi většími úseky textu. Anaforické odkazování mezi nekoreferenčními entitami (ANAF) je velice zajímavý lingvistický jev, který jsme však začali označovat až v poslední fázi zpracování dané práce, a proto výsledky o jeho výskytu v textech zatím nejsou dostatečně reprezentativní.

Vypracovanou klasifikaci jsme aplikovali na koreferenční a anaforické vztahy v Pražském závislostním korpusu. Byla provedena anotace těchto vztahů na polovině korpusu PDT. Dva anotátoři podrobili rozboru cca 25,000 českých vět. Srovnání shody mezi anotátory při navazování vztahů a určování jejich typů ukázalo, že použitá klasifikace je poměrně spolehlivá. Ačkoliv mezianotátorská shoda ještě nedosahuje hodnot, které se považují za dostatečné pro použití anotovaného korpusu pro aplikační účely (F1-measure pro textovou koreferenci se stejně označenými typy nepřesahuje 65%), je ale již cenná pro řešení veškerých lingvistických úloh. Relativně velké rozdíly mezi anotátory jsou způsobeny především ambiguitou koreferenčních vztahů v samotném textu a mohou být omezeny pouze všeobecným zobecňováním a zjednodušením anotačních pravidel, což by vedlo k výraznému ochuzení (nejenom) lingvistické hodnoty anotace. Co se týče asociační anafory, skutečná situace asociačních anaforických vztahů v textu je natolik složitá, že nemůže být v plné míře zohledněna v anotaci. Vypracované anotační schéma asociační anafory představuje výrazné zjednodušení vztahů v textu, text ani nemůže poskytovat jejich jednoznačnou interpretaci, čímž neshoda mezi anotátory také roste.

Problematickým bodem při anotaci koreferenčních vztahů na skutečných souvislých textech je velká referenční a koreferenční ambiguita. Vycházeli jsme z předpokladu, že souvislému gramaticky a logicky správně uspořádanému textu budou za podobných podmínek (úroveň ovládaní jazyka, věk, vzdělání, pohlaví, čitelnost textu aj.) dva adresáti rozumět stejně. V některých případech sice není vyloučena víceznačnost, ale nepočítali jsme s tím, že těch příkladů bude tolik. Hlubší analýza textů a srovnání anotací koreferencí mezi různými anotátory však ukázaly, že možností koreferencí je výrazně více, než se zdá na první pohled, taková víceznačnost však nenarušuje celkově stejné porozumění textu. Není tedy vyloučené, že při automatické anotaci zájmenné a textové koreference se vyskytne velký počet koreferenčních párů, které budou označeny za koreferenční, neshodnou se s ruční anotací, avšak nebudou to přitom chyby. Je však ještě více pravděpodobné, že automatická analýza koreference najde velký počet možností koreferenční homonymie, která nebude intuitivní, podobně, jako je to v případě s automatickou syntaktickou analýzou.

Je nutno rovněž podotknout, že všem našim výsledkům a statistikám nelze přikládat všeobecnou platnost a vyvozovat z nich spolehlivé teoretické závěry o podobě koreferenčních a anaforických vztahů v současné češtině. Vzhledem k omezenému rozsahu anotovaného korpusu lze závěry analýzy chápat jako přibližné a orientační údaje. Je možné, že specifika Pražského závislostního korpusu (publicistický styl, novinové články) výsledky analýzy částečně zkresluje. Můžeme s velkou pravděpodobností předpokládat, že jiná volba anotovaných textů, např. z oblasti umělecké literatury, povede k jiným výsledkům. V jistém smyslu může být tato práce chápána jako formulace hypotézy o tom, jak mohou být pojaty koreferenční a jiné anaforické vztahy v textu. Tuto hypotézu pak může ověřit například automatická nebo částečně automatická analýza výrazně většího počtu textů, což se předpokládá jako samozřejmé pokračování a budoucí aplikace této práce.

Je třeba také upozornit na to, že zpracování koreference a asociační anafory těsně souvisí a je v podstatě součástí celkové analýzy diskurzu, která je také plánována a částečně již probíhá na Pražském závislostním korpusu. Tyto jevy v úzkém smyslu již nejsou součástí tektogramatické struktury, i když anotace koreference i diskurzu probíhá právě na tektogramatické rovině. Tektogramatická rovina PDT je poměrně úplná a rozšířená anotace textové koreference a asociační anafory je v podstatě informace, kterou tam přinášíme navíc. Tato informace již přesahuje sémantiku věty a vztahuje se k její pragmatické stránce, referenci celého textu a jeho komponentů ke skutečnosti apod. V současné době se uvažuje o zavedení hlubší roviny, tj. roviny diskurzu, ale její podoba ještě není dostatečně definována. Pokud však taková rovina bude zavedena, bude logické přesunout na ni i naši anotaci.

Na závěr je třeba ještě jednou upozornit na to, že projekt zpracování koreference a asociační anafory stále ještě probíhá. Každý nový dokument a rozbor shody mezi anotátory přináší další problémy, náměty, zajímavé případy atd. Průběžně se mění statistika a poměr typů koreferenčních vztahů, zlepšují se výsledky mezianotátorské shody a zvětšuje se počet anotovaných dat.

V.1. Další otázky a výhledy

Vypracovaná klasifikace koreferenčních a asociačně anaforických vztahů a jejich analýza je v podstatě prvním krokem v zpracování daného tématu na velkém korpusu českých textů. Výsledky této analýzy nyní mohou sloužit jako východisko pro další výzkum, který se může zaměřit jedním z následujících směrů:

1. Aplikace (automatická nebo částečně automatická) již existujícího rozboru na větší jazykový korpus – jeho použití pro strojové učení, automatický překlad, automatickou generaci výrazů v koreferenčních řetězcích, automatické rozpoznávání anafory (anaphora resolution) a koreference (coreference resolution) atd.
2. Automatické zapojení informací, které jsou součástí tektogramatické syntaktické struktury a jejich využití pro vytvoření koreferenční a anaforické struktury textu. Jde např. o vztahy asociační anafory, které vyplývají z významů funktorů odpovídajících uzlů, označení koreference mezi subjektem a jmennou částí přísudku, která vyplývá ze struktury stromu, označení asociační anafory uvnitř jedné věty atd.
3. Lingvistická teorie anaforických vztahů v textu zpracovaná na základě daného korpusu koreference a asociační anafory a s použitím informací, kterými disponuje tektogramatická rovina PDT. Z tektogramatických stromů lze zjistit, zda daná jmenná fráze je použita v anaforické pozici s ukazovacím zájmenem, jiným zájmenem nebo číslovkou, zda se opakuje stejné pojmenování, jeho synonymum nebo pronominalizace. Dále je možné formulovat hypotézy, v jakých pozicích, jak často a proč se používají odpovídající konstrukce.
4. Lingvistická analýza nalezených teoreticky zajímavých případů, jako např. anafora na nevyjádřený antecedent, anafora na příznak, počet atd.
5. Analýza a zapracování rozdílu mezi identifikačními a predikačními větami do koreferenčních vztahů.

6. Sémantická analýza víceznačnosti koreferenčních vztahů v textu, a tedy i textu samotného, která vyplývá z našeho rozboru mezianotátorské shody.
7. Detailizace a upřesnění rozdílu mezi specifickými a nespecifickými jmennými frázemi a výzkum jejich rozdílného chování v anaforických pozicích v textu.

Obecně je možné konstatovat, že mnohé teoretické otázky textové syntaxe, teorie reference a teorie anafory mohou být přesněji a podrobněji vyřešeny pomocí velkého korpusu anaforicky anotovaných textů. Věříme tedy, že tato práce bude užitečná nejenom pro aplikační účely, ale i pro teoretický výzkum.

Summary

The purpose of this thesis is to describe the theoretical basis of annotation of the extended nominal coreference and the bridging anaphora in the Prague Dependency Treebank.

The Prague Dependency Treebank (PDT 2.0) is a large collection of linguistically annotated data and documentation. In PDT 2.0, Czech newspaper texts are annotated using a three-layer annotation scenario. The most abstract (tectogrammatical) layer includes, among other mark-ups, the annotation of coreferential links.

In PDT 2.0, two types of coreference are annotated: grammatical and textual coreference. The grammatical coreference typically occurs within a single sentence, since the antecedent can be derived on the basis of grammatical rules of a given language. It includes relative pronouns, verbs of control, reflexive pronouns, reciprocity and verbal complements. As for textual coreference, it has been restricted up to now to cases in which a demonstrative *this* or an anaphoric pronoun of the 3rd person, also in its zero form, are used. This thesis focuses namely on the next stage of anaphoric annotation, which is being carried out on PDT now. In this stage, the textual coreference is annotated also for non-pronominal and non-zero NPs, and also for some cases of adjectives, adverbs and verbs. Together with this textual coreference, bridging relations of several types are being annotated.

In the thesis, I propose to base the processing of coreference and bridging anaphora on both theoretical background of the reference theory and practical implementation of coreferenctial data on large textual corpora. A theoretical point of view helped me understand many deep liguistical details of the mechanism of reference, anaphora and coreference. Comparison with the existing schemes of coreference annotation helped me restrict high variety of relations to a reasonable amount that can be processed reliably.

Subject to annotation are pairs of coreferring (by bridging anaphora semantically related) expressions, the preceding expression is called antecedent, the subsequent one is called anaphor. It is possible for an expression to be an antecedent for more than one coreferential and/or bridging expressions at the same time. The reverse is true only for bridging relations, i.e. one expression may have more than one bridging antecedent but just one coreferential antecedent. The coreference and bridging relations are to be marked between elements of the following categories: nouns (*Prague – the town*), anaphoric adverbs (*in the town - there*),

numerals (*by 1999 – this year*), verbs if coreferring with NPs (*They tried to teach him to read – The attempt was not successful.*). Adjectives are annotated only if they are coreferential with a named entity, so e.g. we annotate pairs as *German – Germany*. Names and other named entities are all subjects to annotation. A substring of a named entity, however, is not to be annotated if it is not a named entity itself. Thus, for the sequence *The Charles University of Prague... Prague ...* the two instances of NP *Prague* are to be marked coreferential; but in *Institute of Nuclear Research ... nuclear research* the two instances of NP *research* are not to be coreferred. Due to the syntactic structure of tectogrammatical trees, roots of coordinating and appositional structures can technically also serve as antecedents.

Most of the thesis describes the annotation scheme of extended nominal coreference and bridging anaphora.

Extended textual coreference is further subclassified into two types: coreference of NPs with specific reference (coref_text, type 0) and relations between NPs with generic reference (coref_text, type NR). This decision is made on the basis of the expectation, that generic coreferential chains have different anaphoric rules from the specific ones. This group also includes a big number of abstract nouns whose coreference is not quite clear in every particular case. So, the generic type of textual coreference serves as the ambiguity group too.

Textual coreference covers also the cases of endoforic references to the segment of (preceding) text larger than one sentence, or phrase, including also the cases when the antecedent is understood by inference from a broader co-text. The pronominal anaphoras being already annotated in PDT 2.0, we add links, in which the anaphora is expressed by an NP or an adverb.

A specifically marked link for exophora denotes that the referent is "out" of the co-text, it is known only from the situation. In the same way that it was done for segments, the new nominal and adverbial links are added.

By bridging relations, we annotate only those expressions that are non-coreferential and that stand in some conceptual relation to their antecedent. The participation on the text cohesion is considered to be important, so in ambiguous cases, the relations that are important for the text cohesion are annotated.

At present, we consider the following relations to be relevant:

- part-whole (having two directions PART_WHOLE and WHOLE_PART),
- set-subset/element of the set (also two-directional SET_SUB and SUB_SET),
- object-function (FUNCT for e.g. *class-teacher*),
- CONTRAST for coherence relevant discourse opposites (e.g. *People don't chew, it's cows who chew*),
- ANAF for non-cospecifying anaphoric Nps
- underspecified group REST for capturing bridging references – potential candidates for a new group of bridging relations (e.g. location – resident, relations between relatives (*mother – son*, etc.), event – argument (*listening – listener*) and some other relations).

In some cases, the distinction between SUB_SET and PART groups is quite problematic, so that the only reason to decide for the type of a bridging relation is the countability of corresponding nouns. For the time being, the instruction for such type of ambiguities is to annotate type PART only in clear cases of non-separable parts.

In order to develop maximally consistent annotation scheme, we follow a number of basic principles. Some of them are presented below:

- **Chain principle**: Coreference relations in text are organized in ordered chains. The most recent mention of an entity is marked as antecedent. This principle is checked automatically. The chain principle does not concern bridging anaphora.
- **Principle of the maximum length of coreferential chains.** This principle, similar to the chain principle, concerns only the cases of textual coreference. It states that in case of multiple choices, we prefer to continue the existing coreference chain, rather than to begin a new one. To fulfill this principle, grammatical coreferential chains (already annotated in PDT) are being continued by textual ones, and similarly, the already annotated textual coreferential chains are continued by currently annotated non-pronominal links in turn.

- **Principle of maximal size of an anaphoric expression**. This principle claims that the whole subtree of the antecedent/anaphor is always subject to annotation. This principle is partially governed by the dependency structure of the tectogrammatical trees and may be sometimes counter-intuitive.

- **Principle of cooperation with the syntactic structure of the given dependency tree**. We do not annotate relations that are already captured by the syntactic structure of the tectogrammatical tree. So, for example, we do not annotate predication and apposition relations. Also, bridging relations are not to be annotated if the anaphora is a direct child of its antecedent in the tectogrammatical tree, and it has some of the predefined labels for the valence relations (functors), such as PAT(iens), AUTH(or), APP(urtenance), etc.. So, for example, the relation between *strop* (*ceiling*) and *místnost* (*room*) in the phrase *strop této místnosti* (*the ceiling of this room*) is not annotated, as in the tectogrammatical tree, the node *místnost* has the functor APP, being the direct child of the node *strop*.

- **Principle of primary coreference to anaphora**. Coreference, not anaphora, is subject to textual coreference annotation. Unlike most existing coreference schemes, we try to strictly distinguish identity relations and anaphoric relations. In many cases, an anaphoric relation is also a coreferential relation, although this is not always the case. In a Slavonic language, lacking the grammatical category of definiteness, we cannot afford to choose only definite NPs for anaphoric annotation, so we have to annotate all NPs that refer to the same entity. Non-coreferential anaphoric entities are annotated separately as a bridging relation.

- **Preference of coreference over bridging anaphora**. The preference says that in case of multiple choice, we always prefer textual coreference to bridging relation.

Coreference and bridging annotation is beingperformed using the TrEd annotation tool, developed at the Institute of Formal and Applied Linguistics at Charles University in Prague. The annotation is carried out on tectogrammatical tree structures assigned to the sentences in text. The present scenario of PDT provides a number of coreferential attributes. Coreference relations are captured by arrows leading from the anaphor to the antecedent and the various types of relations (bridging, textual, grammatical) are distinguished by different colours of the arrows.

The annotation scheme described in the thesis has been applied on a large scale to the whole PDT corpus by two instructed annotators, students of linguistics. So far, 50% of PDT has been annotated.

For the purpose of checking and improving the annotation guidelines, we regularly provide and describe the inter-annotator measurements. A detailed study of the texts annotated by both annotators revealed several sources of typical errors. The inter-annotator agreement is also greatly affected by parameters of the text as a whole. The interpretations of short texts are generally far less than of the longer texts of 20 to 120 sentences. Agreement is getting more difficult, the more complex the judgments that the annotators have to make become. Also, the degree of abstraction plays a crucial role in the results of the inter-annotator agreement.

The first phase of the coreference annotation process has revealed several problematic cases concerning annotation of anaphoric relations in Czech. The most problematic aspect in annotating textual coreference concerns abstract nouns. Given that in some cases such NPs are clearly coreferential and anaphoric, we cannot exclude them from the annotation. However, there are many more cases in which the decision for postulation of coreference is not certain, sometimes appearing to be quite redundant. The following questions arise when annotation of abstract nouns is carried out: Should we annotate such cases at all? If we annotate them, what kind of coreference type is that (specific or non-specific coreference)? For the time being, we annotate relations between abstract nouns as generic coreference (coref_text, type 0), in order to be able to exclude them if needed. Yet, there still remains the problem of distinguishing between abstract and concrete nouns, the boundary between them being rather gradual.

There are some other questions left unanswered, such as annotating coreference in prepositional phrases, annotation of complex nouns, etc., which are mainly solved using formal conventions.

Seznam zkratek a značek

ACE – Automatic Content Extraction

bridging – asociační anafora

coref_gram – gramatická koreference

coref_text – textová koreference

FUNCT – asociační anafora typu „entita – funkce“

MUC – Message Understanding Conferences

NE – named entity (pojmenovaná entita)

NLP – Natural Language Processing

NP – jmenná fráze

PART – asociační anafora typu „část – celek“

PDT – The Prague Dependency Treebank

PoCoS – Postdam Commentary Corpus

PP – předložková fráze

SUBSET – asociační anafora typu „podmnožina – množina“

TFA – Topic-Focus Articulation

t-lemma – tektogramatické lemma

TGS – tektogramatická struktura

TrEd – Tree Editor

ÚFAL – Ústav formální a aplikované lingvistiky na MFF UK

UZ – ukazovací zájmeno

VL – vymyšlený vlastní příklad

Zkratky funktorů tektogramatické roviny, které jsou použity v práci:

ACMP (od *accompaniment*) – funktor pro takové volné doplnění, které vyjadřuje způsob uvedením nějaké okolnosti;

ACT (od *actor*) – funktor pro první aktant;

ADVS (od *adversative*) – funktor pro kořen takové souřadné struktury, která reprezentuje koordinační spojení, v němž jsou spojeny zpravidla dva obsahy, které nejsou v souladu; v pořadí druhý obsah je v rozporu s očekáváním plynoucím z obsahu prvního;

APP (od *appurtenance*) – funktor pro volné doplnění substantiv označující osobu nebo věc, ke které je osoba nebo věc vyjádřená řídícím substantivem ve vztahu přináležitosti.

APPS (od *apposition*) – funktor pro kořen takové souřadné struktury, která reprezentuje apoziční spojení;

AUTH (od *author*) – funktor pro volné doplnění substantiv, které označuje tvůrce, autora artefaktů.

CPHR (od *compound phraseme*) – funktor pro jmennou část složených predikátů a pro neslovesnou část kvazimodálních sloves tvořených slovesem *být* a predikativním adverbiem;

CONFR (od *confrontation*) – funktor pro kořen takové souřadné struktury, která reprezentuje koordinační spojení, ve kterém se zpravidla dva rozdílné nebo přímo kontrastní obsahy stavějí proti sobě, vzájemně se konfrontují;

CONJ (od *conjuction*) – funktor pro kořen takové souřadné struktury, která reprezentuje koordinační spojení vyjadřující prosté slučování dvou a více obsahů;

ID (od *identity*) – funktor pro efektivní kořen identifikačního výrazu, který zachycujeme jako identifikační strukturu;

MAT (od *material, partitiv*) – funktor pro aktant substantiv, který označuje obsah (osoby, věci, látku, materiál aj.) kontejneru vyjádřeného řídícím substantivem;

PAT (od *patiens*) – funktor pro druhý aktant;

PREC (od *reference to preceding text*) – funktor pro uzel, který reprezentuje výraz signalizují návaznost klauze na předcházející kontext;

RSTR – funktor pro volné doplnění, které blíže vymezuje řídící substantivum.

Zástupná tektogramatická t-lemmata, které jsou použity v práci:

#Bracket – t-lema uzlu reprezentujícího symbol závorky „ (“ nebo „) “;

#Colon – t-lema uzlu reprezentujícího symbol dvojtečky „ : “ ;

#Comma – t-lema uzlu reprezentujícího interpunkční čárku „ , “;

#Cor – t-lema nově vytvořeného uzlu zastupujícího v povrchové podobě věty zpravidla nevyjádřitelný kontrolovaný člen v konstrukcích s kontrolou

#Dash – t-lema uzlu reprezentujícího symbol pomlčky nebo spojovníku;

#Forn – t-lema nově vytvořeného uzlu vystupujícího jako řídící uzel cizojazyčného výrazu; uzel s tímto t-lematem nemá v povrchové podobě věty protějšek

#Gen – t-lema nově vytvořeného uzlu zastupujícího v povrchové podobě věty nepřítomný všeobecný aktant;

#Idph – t-lema nově vytvořeného uzlu, který slouží jako pomocný uzel pro zachycení identifkačních výrazů;

#Percnt – t-lema uzlu reprezentujícího symbol procenta „ % “;

#PersPron – t-lema uzlu reprezentujícího osobní nebo posesivní zájmeno (včetně zájmen refexivních), a to jak u uzlů nově vytvořených, tak u uzlů reprezentujících povrchově realizované zájmeno. U uzlů nově vytvořených signalizuje t-lema #PersPron aktuální elipsu.

#Qcor – t-lema nově vytvořeného uzlu zastupujícího v povrchové podobě věty zpravidla nevyjádřitelné valenční doplnění v konstrukcích s kvazikontrolou;

#Rcp – t-lema nově vytvořeného uzlu zastupujícího valenční doplnění, které v povrchové podobě věty není přítomno z důvodu reciprokalizace;

#Unsp – t-lema nově vytvořeného uzlu zastupujícího v povrchové podobě věty nerealizované blíže nespecifkované valenční doplnění.

Seznam obrázků

Seznam tabulek

Literatura

ADAMEC, Přemysl. K vyjadřování referenční určenosti v češtině a ruštině. *Slovo a slovesnost*, 1980, roč. 41, s. 257–264.

ADAMEC, Přemysl. Funkcii ukazatel'nych mestoimenij v češskom jazyke v sravnenii s russkim. In ŠIROKOVÁ, Alexandra G.; GRABJE, Vladimír. *Sopostavitel'noje izučenije grammatiki i leksiki russkogo jazyka s češskim jazykom i drugimi slavjanskimi jazykami*. Moskva: izdatel'stvo MGU, 1983, s. 173–190.

ADAMEC, Přemysl. Različija v vyraženii anaforičeskich otnošenij meždu russkim i češskim jazykami. *Russkij jazyk za rubežom*. Moskva, 1984. s. 73–78.

ADAMEC, Přemysl. K prostředkům textové syntaxe v současné češtině. In *Přednášky z XXX. Běhu LŠSS, 1986*; UK, Praha, 1988, s.105–115.

ANDERSON, A., M. BADER, E. BARD, E. BOYLE, G. M. DOHERTY, S. GARROD, S. ISARD, J. KOWTKO, J. McALLISTER, J. MILLER, C. SOTILLO, H. S.THOMPSON, and R. WEINERT. The HCRC Map Task Corpus. *Language and Speech*, roč 34, 1991, s. 351–366.

ARUT'UNOVOVÁ (ARUT'UNOVA), Natalie D. *Predloženije i jego smysl*. Moskva: URSS, 1976 (reprint 2005).

AVERINTSEVA-KLISCH (AVERINTSEVOVÁ-KLISCHOVÁ), Maria; CONSTEN, Manfred. The role of discourse topic and proximity for demonstratives in German and Russian. In BERGLJOT, Behrens; FABRICIUS-HANSEN, Cathrine; HASSELGÅRD, Hilde; JOHANSSON, Stig (eds.). *Information Structuring Resources in Contrast*. Amsterdam: John Benjamins, 2007, s. 221–240.

BEL'SKIJ, Andrej V. Intonacija kak sredstvo determinirovanija i predecirovanija v russkom literaturnom jazyke. In SUCHOTIN, Vladimir P. (ed.). *Issledovanija po sintaksisu russkogo literaturnogo jazyka. Sbornik statij*. Moskva: Akadamija nauk SSSR, 1956, s. 188–199.

BENACCHIO, Richard. K voprosu ob opredelennom artikle v slavjanskich jazykach: rez'janskij govor. In DULIČENKO, Alexandr D. (ed.). *Jazyki malye i bol'šie*. Tartu: Slavica Tartuensia, 1998, s. 76–88.

BIRKENMAIER, Willy. *Artikelfunktionen einer artikellosen Sprache. Studien zur nominalen Determination im Russischen*. München: Wilhelm Fink Verlag,1979.

BOGOCZOVÁ, Irena. Specifické funkce zájmena *ten* mluvených komunikátech. In *Tváře češtiny.* Ostrava: FF Ostravské univerzity, 2000, s. 112–119.

BOGUSLAVSKAJA, Olga Ju.; MURAV'EVA, Irina A. Mechanizmy anaforičeskoj nominacii. In KIBRIK Alexandr E.; NARIN'JANI Alexandr S. (eds.). *Modelirovanie jazykovoj dejatel'nosti v intellektual'nych sistemach.* Moskva: Nauka, 1987, s. 78–128.

BRUNHUBER, Brigitte. Aspekt und Determiniertheit im Russischen. *Die slawischen Sprachen* roč. 3, 1983, s. 5–13.

CARLSON, Lynn; MARCU, Daniel; OKUROWSKI, Mary Ellen. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In KUPPEVELT, Jan van; SMITH Ronnie (eds.). *Current Directions in Discourse and Dialogue.* Kluwer: Academic Publishers, 2003, s. 85–112.

CHAMBERLAIN, Jon; POESIO, Massimo; KRUSCHWITZ, Udo. Phrase Detectives: A Webbased Collaborative Annotation Game. In AUER, Sören; SCHAFFERT, Sebastian, PELLEGRINI, Tassilo (eds.). *Proceedings of the International Conference on Semantic Systems (I-Semantics'08).* 2008a.

CHAMBERLAIN, Jon; POESIO, Massimo; KRUSCHWITZ, Udo. Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In RAIKO, Tapani; HAIKONEN, Pentti; VAYRYNEN, Jaakko. *Proceedings of STEP2008*, Venice: Chamberlain. 2008b.

CINKOVÁ, Silvie. Semantic Representation of Non-Sentential Utterances in Dialog. In *Proceedings of SRSL 2009, the 2nd Workshop on Semantic Representation of Spoken Language.* Association for Computational Linguistics, Athina, Greece, 2009, s. 26–33.

CLARK, Herbert H. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, June 10–13, Cambridge, Massachusetts, 1975.

COHEN, Jacob. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, roč. 20(1), 1960, s. 37–46.

CONSTEN, Manfred; KNEES, Mareile; SCHWARZ-FRIESEL(OVÁ), Monika. The function of complex anaphors in Text. In SCHWARZ-FRIESEL(OVÁ), Monika, CONSTEN, Manfred; KNEES, Mareile (eds.). *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference.* Amsterdam: John Benjamins B.V. 2007., s. 81–102.

COLLINS, Michael; SINGER, Yoram. Unsupervised Models for Named Entity Classication. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC).* 1999, s. 189–196.

CORBETT, Greville G. The use of the genitive or accusative for the direct object of negated verbs in Russian: a bibliography. In BRECHT, Richard D; LEVINE, James S. (eds.), *Case in Slavic*. Columbus: Ohio, 1986, s. 361–372.

CORNISH, Francis. Indirect pronominal anaphora in English and French. In SCHWARZ-FRIESEL(OVÁ), Monika, CONSTEN, Manfred; KNEES, Mareile (eds.). *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference.* Amsterdam: John Benjamins B.V. 2007., s. 21–36.

ČERNĚJKO, Ludmila O. *Abstraktnoje imja. Lingvo-filosofskij analis abstraktnogo imeni.* Moskva: MGU, 1997.

DAHL, Östen. On Generics. In KEENAN, Edward (ed.), *Formal Semantics of Natural Language. Cambrigde*, London & New York, 1975, s. 99–112.

DANEŠ, František. O identifikaci známé (kontextově zapojené) informace v textu. *Slovo a slovesnost*, 1979, roč. 40, s. 257–270.

DODDINGTON, George; MITCHELL, Alexis; PRZYBOCKI, Mark; RAMSHAW, Lance; STRASSEL, Stephanie; WEISCHEDEL, Ralph. The Automatic Content Extraction (ACE) program – Tasks, data, and evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, 2004.

DONNELLAN, Kieth S. Reference and definite description. *Philiosophical Review*, 1966, roč. 75, s. 281–304.

DONNELLAN, Kieth S. Speaker reference, descriptions and anaphora. In FRENCH, Peter; UEHLING, T.E. Jr; WETTSTEIN, H.K. *Contemporary perspectives in the Philosophy of Language*. Minneapolis: U. of Minnesota Press, 1979, p. 28–44.

ERKÜ, Feride; GUNDEL, Jeanette.K. The pragmatics of indirect anaphors. In VERSCHUEREN, Jef; BERTUCCELLI PAPI, Marcela (eds.). *The Pragmatic perspective: Selected Papers from the 1985 International Pragmatics Conference.* Amsterdam: John Benjamins.1987, s. 533–545.

FAUCONNIER, Gilles. *Mental spaces. Aspects of meaning construction in natural languages.* Cambridge: Cambridge University Press. 1995.

FREGE, Gottlob. Über Begriff und Gegenstand. *Vierteljahresschrift für wissenschaftliche Philosophie*. Leipzig, roč. 16, 1892, s. 192–205.

GAK, Vladimir G. *Sopostavitelnaja leksikologija*. Moskva: Vyššaja škola, 1977.

GARDENT, Claire, MANUELIAN, Helene, KOW, Eric..Which bridges for bridging definite descriptions? In *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, Budapest, 2003, s. 69–76.

GIVON, Talmy. The grammar of referential coherence as mental processing instructions. *Linguistics* roč. 30, 1992, s. 5–55.

GLADROW, Wolfgang. *Die Determination des Substantivs im Russischen und Deutschen.* Leipzig:Verlag Enzyklopädie Leipzig.1979.

GLADROW, Wolfgang. Semantika i vyraženie opredelennosti/neopredelennosti. In *Teorija funkcional'noj grammatiki IV. Subjektnost'. Objektnost'. Kommunikativnaja perspektiva vyskazyvanija. Opredelennost'/neopredelennost'.* Sankt-Peterburg: Nauka, 1992, s. 232–266.

GOLOVAČEVA, Anna V. Identifikacja i individualizacii v anaforičeskich strukturach. In NIKOLAEVA, Tatiana M. (ed.) *Kategorija opredelennost/neopredelennosti v slavjanskix i balkanskix jazykach.* Moskva: Nauka, 1979, s. 175–203.

HAIZHOU, Li; KUMARAN, A. *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration.* Suntec, Singapore. 2009

HAWKINS, John A. *Definiteness and Indefiniteness: A study in reference and grammaticality prediction.* London: Groom Helm, 1978.

HEIM, Irene. Artikel und Definitheit. In STECHOW, Arnim von; WUNDERLICH, Diter (eds). *Semantik. Ein Internationales Handbuch der zeitgenössischen Forschung. Berlin: Walter de Gruyter,* 1991, s. 487–535.

HELBIG, Hermann. *Knowledge Representation and the Semantics of Natural Language.* Berlin: Springer-Verlag. 2006.

HENSCHEL Renate; CHENG, Hua; POESIO, Massimo. Pronominalization revisited. In KAUFMANN, Morgan (ed.). *Proceedings of 18th COLING.* Saarbrücken: Universität des Saarlandes, 2000, s.306–312.

HIRSCHMAN, Lynette. MUC-7 coreference task definition version 3.0. In CHINCHOR, Nancy (ed.) *Proceedings of the 7th Message Understanding Conference.*1997.

HLAVSA, Zdenek. K protikladu určenosti v češtině. *Slovo a slovesnost,* roč. 33, 1972, s.199–203.

HLAVSA, Zdenek Palkova kniha o mezivětném odkazování, *Slovo a slovesnost,* roč. 33, 1972, s.47–52.

HLAVSA, Zdenek. *Denotace objektu a její přostředky v současné češtině.* Praha: Academia, 1975.

HRBÁČEK Josef . *Nárys textové syntaxe spisovné češtiny.* Praha: Trizonia, 1994.

CHENG, Hua; POESIO, Massimo; HENSCHEL, Renate; MELLISH, Chris. Corpus-based NP modifier generation. In *Proceedings of the Second NAACL.* Pittsburgh, 2001.

CHIARCOS, Christian; KRASAVINA, Olga. PoCoS – Potsdam Coreference Scheme. In *Proceedings of ACL-2007 Linguistic Annotation Workshop*. Praha, 2007, s. 156–163.

KABADJOV, Mijail A.; POESIO, Massimo; STEINBERGER, Josef. Task-Based Evaluation of Anaphora Resolution: The Case of Summarization. In *Proceedings of RANLP Workshop on Recent Developments in Summarization*, Varna, Bulgaria, 2005.

KATZ, Jerrold J. The neoclassical theory of reference In FRENCH, P.A.; UEHLING, T.F., WETTSTEIN, H. K. (eds.) *Contemporary perspectives in the philosophy of language*. Minneapolis: University of Minnesota Press, 1979, s. 103–124.

KOSESKA-TOSZEWA, Violetta. O kategorii określoności – nieokreśloności w planie konfrontatywnym na przykładzie z języka bułgarskiego, polskiego i rosyjskiego. *Z polskich studiów slawistycznych*, seria VI. Warszawa: Językoznawstwo, 1983, s. 187–194.

KRAVALOVÁ, Jana; ŽABOKRTSKÝ, Zdeněk. Czech Named Entity Corpus and SVM-based Recognizer. In *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 194–201, Suntec, Singapore, 7 August 2009*. 2009, s. 194–201.

KREJDLIN, Grigory E.; RACHILINA, Ekaterina V. Denotativnyj status otglagol'nych imen *Naučno-techničeskaja informacija*, seria 2, roč. 12, 1981, s. 17–22.

KOMÁREK, Miroslav. Sémantická struktura deiktických slov v češtině. *Slovo a slovesnost*, roč.39, 1978, s.5–14.

KŘÍŽKOVÁ, Helena. Zájmena typu *ten* a *takový* v současných slovanských jazycích. *Slavica Slovaca*, roč. 6, 1971, č.1, s.15–30.

KUČOVÁ, Lucie; KOLÁŘOVÁ, Veronika; ŽABOKRTSKÝ, Zdeněk; PAJAS, Petr, ČULO, Oliver. *Anotování koreference v Pražském závislostním korpusu*. Praha: UFAL/CKL MFF UK, 51, Technická zpráva-2003-19, 2003.

KUČOVÁ, Lucie; HAJIČOVÁ, Eva. Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of 5th Discourse Anaphora and Anaphor Resolution Colloquium*. Edicoes Colibri, 2004.

LAVRIC, Eva. *Fülle und Klarheit*. Eine Determinantensemantik. Deutsch – Französisch – Spanisch. Band I: Referenzmodell. Band II: Kontrastiv-semantische Analysen. Tübingen: Stauffenburg Linguistik. 2001.

LENZ, Friedrich. Reflexivity and temporality in discourse deixis. In SCHWARZ-FRIESEL(OVÁ), Monika, CONSTEN, Manfred; KNEES, Mareile (eds.). *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference*. Amsterdam: John Benjamins B.V., 2007, s.69–80.

MAES, Alfons. Referent ontology and centering in discourse. *Journal of semantics*. roč. 14, 1997, s. 207–235.

MAKHOUL, John; KUBALA, Francis; SCHWARTZ, Richard; WEISCHEDEL, Ralph. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*. Herndon, VA, February 1999.

MARX, Konstanze; BORNKESSEL(OVÁ)-SCHLESEWSKY, Ina; SCHLESEWSKY, Matthias. Resolving complex anaphors. In SCHWARZ-FRIESEL(OVÁ), Monika, CONSTEN, Manfred; KNEES, Mareile (eds.). *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference*. Amsterdam: John Benjamins B.V., 2007, s. 259–277.

MATHESIUS, Vilém. Přívlastkové *ten, ta, to* v hovorové češtině. *Naše řeč*, roč.10, 1926, s.39–41.

MEL'ČUK, Igor. A. *Opyt teorii lingvističeskich modeley „Smysl ↔ Text"*. Moskva: Nauka, 1974 (2. vydání 1999).

MIKULOVÁ, Marie. a kol. *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka, I, II.* Technická zpráva ÚFAL TR-2005-28. Praha: Universitas Carolina Pragensis, 2005.

MILLER, George A.; BECKWITH, Richard; FELLBAUM, Christiane; GROSS, Derek; MILLER, Katherine. Five papers in WordNet. In FELLBAUM, Christiane (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

MUÑOZ, Rafael; SAIZ-NOEDA, Maximilian; SUÁREZ, Armando; PALOMAR, Manual. Semantic approach to Bridging Reference Resolution. In *Proceedings of Machine Translation 2000*. Exeter (UK): University of Exeter. 2000.

MITKOV, Ruslan. *Anaphora Resolution*. UK: Longman, 2002.

MLADOVÁ, Lucie; ZIKÁNOVÁ, Šárka; HAJIČOVÁ, Eva. From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*. Marrakech, Morokko, 2008.

MLADOVÁ, Lucie. *Diskurzní vztahy v češtině a jejich zachycení v anotovaném korpusu*. Nepublikovaná diplomová práce. Praha: Filozofická fakulta Univerzity Karlovy, 2008.

MÜLLER, Christoph; STUBE, Michael. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark, 2001, s. 90–95.

NEDOLUZHKO, Anna. Ukazovací zájmeno *ten* a generické jmenné fráze v češtině. In *IV. mezinárodní setkání mladých lingvistů Olomouc 2003: Jazyky v kontaktu, jazyky v konfliktu*. Olomouc: Univerzita Palackého v Olomouci, 2003, s. 85 – 96.

NEDOLUZHKO, Anna. Takový? Ten? Takový ten. Konkurence a význam. In *Opera Academiae Paedagogicae Liberecensis. Series Bohemistica vol. III. Eurolingua 2004.* Liberec: TUL, 2005, s. 92–105.

NEDOLUZHKO, Anna. Ukazovací zájmeno *ten* v kontextu dnešních bádání. In ULIČNÝ, Oldřich (ed.). *Opera Linguae Bohemicae Studentium 7, Úvahy o jazyce a literatuře.* Praha: Filozofická fakulta univerzity Karlovy, 2005, s. 11 – 24.

NEDOLUZHKO, Anna; MÍROVSKÝ, Jiří; PAJAS, Petr. The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of ACL-IJCNLP 2009, Linguistic Annotation Workshop (LAW III)*. Suntec, Singapore, 2009.

NEDOLUZHKO, Anna. Razmetka koreferencii na sintaksičeski annotorovannom korpuse češskich tekstov. In KIBRIK, Alexandr E. a kol. (eds.). *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2009"*. Issue 8 (15). 2009, Moskva: RGGU, s. 332 – 339.

NOVAK, Vaclav; HARTRUMPF, Sven; HALL, Keith. Large-scale Semantic Networks: Annotation and Evaluation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions.* Boulder, USA, 2009.

NOVÁK, Václav; HALL, Keith. Inter-sentential Coreferences in Semantic Networks: An Evaluation of Manual Annotation. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).* Marrakech, Morocco, 2008.

ORASAN, Constantin. PALinkA: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue.* Sapporo, 2003.

PADUČEVOVÁ (PADUČEVA), Elena V. Nacional'nyj korpus russkogo jazyka kak resurs při issledovanii predmetnoj sootnesennosti imen. In *Konference NTI-2007. Vserossijskij institut naučnoj i techničeskoj informacii (VINITI AV RF).* 2007.

PADUČEVOVÁ (PADUČEVA), Elena V. Denotativnyj status imennoj gruppy i ego otraženie v semantičeskom predstavlenii predloženija. *Naučno-techničeskaja informacija,* ser. 2, roč. 9, 1979, s. 25–31.

PADUČEVOVÁ (PADUČEVA), Elena V. K teorii referencii: imena i deskripcii v neèkstensional'nych kontekstach. *Naučno-techničeskaja informacija*, ser. 2, roč. 1, 1983, s. 24–29.

PADUČEVOVÁ (PADUČEVA), Elena V. *Vyskazyvanije i jego sootnesennost' s dejstviteľnostju.* Moskva: Nauka, 1985.

PADUČEVOVÁ (PADUČEVA), Elena V. O referencii jazykovych vyraženij s nepredmetnym značenijem. *Naučno-techničeskaja informacija*, cep.2, roč. 1, 1986.

PADUČEVOVÁ (PADUČEVA), Elena V. Predloženija toždestva: Semantika i kommunikativnaja struktura. In PETROV, Vladimir V. (ed.) *Jazyk i logičeskaja teorija.* Moskva: Nauka, 1987, s. 152–163.

PADUČEVOVÁ (PADUČEVA), Elena V. Snova anafora i koreferentnost'. In: USPENSKIJ, Boris (ed.). *Voprosy kibernetiki. Problemy razrabotki formal'noj modeli jazyka.* Moskva: Nauka, 1988, s. 71–88.

PAJAS, Petr; ŠTĚPÁNEK, Jan. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the The 22nd Interntional Conference on Computational Linguistics*. Manchester, 2008, s. 673–680.

PALEK, Bohumil. *Cross-reference: a study from hyper-syntax.* Praha: Filozofická fakulta Univerzity Karlovy, 1968.

PALEK, Bohumil. *Referenční výstavba textu.* Praha: Univerzita Karlova, 1988.

PASSONNEAU, Rebecca. *Instructions for applying Discourse Reference Annotation for Multiple Applications (DRAMA).* Nepublikovaný rukopis, 1996.

POESIO, Massimo; CHENG, Hua; HENSCHEL, Renate; HITZEMAN, Janet; KIBBLE, Rodger; STEVENSON, Rosemary. Specifying the Parameters of Centering Theory: a Corpus-Based Evaluation using Text from Application-Oriented Domains. In *Proceedings of the 38th ACL.* Hong Kong, 2000b.

POESIO, Massimo. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC – 2000).* Atény, květen 2000c, s. 211–218.

POESIO, Massimo. Associative descriptions and salience: a preliminary investigation. In *Proceedings of the ACL Workshop on Anaphora.* Budapest, duben, 2003.

POESIO, Massimo. The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL.* Boston, duben, 2004a.

POESIO, Massimo. Discourse Annotation and Semantic Annotation in the GNOME Corpus. In *Proceedings of the ACL – 2004 Workshop on Discourse Annotation.* 2004c, s. 72–79.

POESIO, Massimo; BRUNESEAUX, Florence; ROMARY, Laurent. The MATE meta-scheme for coreference in dialogues in multiple language. In *Proceedings of the ACL Workshop on Standards for Discourse Tagging.* Maryland, červen, 1999.

POESIO, Massimo; MEHTA, Rahul; MAROUDAS, Axel; HITZEMAN, Janet. Learning to resolve bridging references. In *Proceedings of ACL.* Barcelona, červenec, 2004b.

POESIO, Massimo; URYUPINA, Olga; VIEIRA, Renata; ALEXANDROV-KABADJOV, Mijail; GOULART, Rodrigo. Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *Proceedings of the ACL Workshop on Reference Resolution.* Barcelona, červenec, 2004c.

POESIO, Massimo; ALEXANDROV-KABADJOV, Mijail. A general-purpose, off-the-shelf system for anaphora resolution. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC – 2004).* Lisbon, květen, 2004.

POESIO, Massimo; MODJESKA, Natalia N. The THIS-NPs Hypothesis: A Corpus-Based Investigation. In *Proceedings of DAARC.* Lisbon, září, 2002.

POESIO, Massimo; NISSIM, Malvina. Salience and possessive NPs: the effects of animacy and pronominalization. In *Proceedings of AMLAP.* Saarbruecken, září, 2001.

POESIO, Massimo; VIEIRA, Renata. A Corpus-based Investigation of Definite Description Use. *Computational Linguistics*, roč. 24, č. 2, 1998, s. 183–216.

POESIO, Massimo; ARSTEIN, Ron. Anaphoric Annotation in the ARRAU Corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC2008).* Marrakech, Morocco, květen, 2008.

POSPELOV, Nikolaj S. O sintaksičeskom vyrazenii kategorii opredelennosti – neopredelennosti v sovremennom russkom jazyke. In: POSPELOV, Nikolaj S. (ed.). *Issledovanija po sovremennomu russkomu jazyku.* Moskva: izdateľstvo MGU, 1970, s. 182–189.

PRASAD, Rashmi; DINESH, Nikhil; LEE, Alan; MILTSAKAKI, Eleni; ROBALDO,Livio; JOSHI, Aravind; WEBBER, Bonnie. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).* Marrakech, Morocco, květen, 2008.

QUINE, Willard van Orman. *Word and Object.* Cambridge: MIT Press, 1960.

RACHILINA Ekaterina; KREJDLIN Grigory E.: Denotativnyj status otglagoľnych imen. *Naučno-techničeskaja informacija,* ser. 2, roč. 12, 1981, s.17–22.

RECASENS(OVÁ), Marta; MARTÍ, Antònia; TAULÉ, Mariona. Text as Scene: Discourse Deixis and Bridging Relations. *Procesamiento del Lenguaje Natural.* Sevilla, Spain, č. 39, 2007, s. 205–212.

RECASENS(OVÁ), Marta; MARTÍ, Antònia. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. In *Language Resources and Evaluation.* 2010.

RUSSELL, Bertrand. On denoting. *Mind,* roč. 14, 1905, s. 479–493.

SANTOS, Diana; SECO, Nuno; CARDOSO, Nuno; VILELA, Rui. HAREM. An Advanced NER Evaluation Contest for Portuguese. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC).* Genoa, Itálie, 2006, s. 1986–1991.

SASSANO, Manabu; UTSURO, Takehito. Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition. In KAUFMANN, Morgan (ed.). *Proceedings of the 18th International Conference on Computational Linguistics (COLING). Volume II.* San Fransisco, 2000, s. 705–711.

SCHNEIDEROVÁ, Eva. K užívání zájmena *ten* (v přívlastkové pozici) v mluvených projevech. *Naše řeč*, roč. 76, 1993, s.31–37.

SCHWARZ, Monika. Textuelle Progression durch Anaphern. Aspekte einer prozeduralen Thema – Rhema Analyze. *Linguistische Arbeitsberichte*, roč. 74, 2000, s. 111–126.

SCHWARZ-FRIESEL(OVÁ), Monika; CONSTEN, Manfred; KNEES, Mareile (eds.). *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference.* Amsterdam: John Benjamins B.V., 2007.

SCHWARZ-FRIESEL(OVÁ), Monika. Indirect Anaphora in text. A cognitive account. In SCHWARZ-FRIESEL(OVÁ), Monika; CONSTEN, Manfred; KNEES, Mareile (eds.). *Anaphors in Text. Cognitive, formal and applied approaches to anaphoric reference.* Amsterdam: John Benjamins B.V., 2007, s.3–20.

SEARLE, John R. *Speech acts: An essay in the philosophy of language.* UK: Cambridge university press. 1969.

ŠEVČÍKOVÁ Magda; ŽABOKRTSKÝ, Zdeněk; KRŮZA, Oldřich. Named Entities in Czech: Annotating Data and Developing NE Tagger. In *Lecture Notes In Computer Science: Proceedings of the 10th International Conference on Text, Speech and Dialogue.* Plzeň: Springer, 2007a, s. 188–195.

ŠEVČÍKOVÁ Magda; ŽABOKRTSKÝ, Zdeněk; KRŮZA, Oldřich. *Zpracováaní pojmenovaných entit v českých textech.* Technická zpráva. Praha: ÚFAL MFF UK, 2007b.

SLEZÁKOVÁ, Markéta. Role ukazovacích zájmen *ten, ta, to* v mluveném dialogickém textu. In *Komunikační a strukturní aspekty češtiny a jiných jazyků*, Praha: FF UK, 1999, s.77–90.

ŠMELEV, Alexey D. *Opredelennost' – neopredelennost' v nazvanijach lic v russkom jazyke.* nepublikovaná dizertační práce. Moskva, 1984.

ŠMELEV, Alexey D. *Referencial'nye mechanizmy russkogo jazyka.* Tampere: Slavica Tamperensia, 1996.

STEPANOV, Jury S. *Imena, predikaty, predloženija (semiologičeskaja grammatika).* Moskva: Editorial URSS, 2004.

STEDE, Manfred. The Potsdam Commentary Corpus. In *Proceedings of the ACL – 2004. Workshop on Discourse Annotation.* Barcelona, 2004, s. 96–102.

ŠTÍCHA, František. K deikticko-anaforickým funkcím lexému *ten. Slovo a slovesnost*, roč. 60, 1999, s.123–135.

TALUKDAR, Partha Pratim; BRANTS, Thorsten; LIBERMAN, Mark; PEREIRA, Fernando. A Context Pattern Induction Method for Named Entity Extraction. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL–X).* New York, červen 2006, s. 141–148.

TUTIN, Agnes; TROUILLEUX, Francois; CLOUZOT, Catherine; GAUSSIER, Eric; ZAENEN, Annie; RAYOT, Stephanie; ANTONIADIS, Georges. Annotating a large corpus with anaphoric links. In *Proceedings of the 3rd Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2000).* Lancaster University, listopad 2000.

UFIMCEVA Anna A. *Tipy slovesnych znakov.* Moskva: Nauka, 1974.

UFIMCEVA Anna A. *Lexičeskoje značenije. Princip semiologičeskogo opisanija leksiki.* Moskva: URSS, 1986.

UHLÍŘOVÁ, Ludmila. Určenost nominální skupiny. In BĚLIČOVÁ, Helena; UHLÍŘOVÁ, Ludmila. *Slovanská věta.* Praha: Euroslavica, 1996, s.225 – 249.

VATER, Heinz. Referenz und Determination im Text. In ROSENGREN, Inger (ed.). *Sprache und Pragmatik, Lunder Symposium 1984 (Lunder Germanische Forschungen 54).* 1984, s. 323–344.

VIEIRA, Renata; Teufel, Simone. Towards resolution of bridging descriptions. In *Proceedings to 35th Annual Meeting of the Association for Computational Linguistics.* Saarbrücken, Germany, 1997.

VIEIRA, Renata; POESIO, Massimo. An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics*, roč. 26, č. 4, 2000, s. 539–593.

WEISS, Daniel. Identitätsaussagen im Russischen: Ein Versuch ihrer Abgrenzung gegenüber anderen Satztypen. In GIRKE, Wolfgang; JACHNOW, Helmut (eds.). Slavistische Linguistik 1977. München, 1978., s. 224–259.

WEISS, Daniel. Indefinite, definite und generische Referenz in artikellosen slavischen Sprachen. In MEHLIG, Hans Robert (ed.). Slavistische Linguistik 1982. München, 1983, s. 229–261.

YOKOYAMA, Olga B. *Kognitivnaja modeľ diskursa i russkij porjadok slov*. Moskva: Jazyki slavjanskoj kultury, 2005.

YULE, George. Pragmatically-controlled anaphora. *Lingua,* roč. 49, 1979, s. 127–135.

ZIKÁNOVÁ, Šárka. *Possibilities of Discourse Annotation in Prague Dependency Treebank (Based on the Penn Discourse Treebank Annotation)*. Technical report. Institute of Formal and Applied Linguistics, Charles University, Prague, 2007.

ZIMOVÁ, Ludmila. *Způsoby vyjadřování větných členů v textu. Konkurence pojmenování, pronominalizace a elize*. Ustí nad Labem: Univerzita Jana Evangelisty Purkyně, 1994.

ZUBATÝ, Josef. Ten. *Naše řeč*, roč. 1, č. 10, 1917, s.253–259.

Internetové odkazy

BERGER, Tilman. *Das System der tschechischen Demonstrativpronomina. Nepublikovaný rukopis*. München 1993. Dostupné na http://homepages.uni-tuebingen.de/tilman.berger/Texte//texte.html

HAJIČ, Jan; HAJIČOVÁ, Eva; HLAVÁČOVÁ, Jaroslava, KLIMEŠ, Vladislav; MÍROVSKÝ, Jiří; PAJAS, Petr; ŠTĚPÁNEK Jan; VIDOVÁ-HLADKÁ, Barbora; ŽABOKRTSKÝ, Zdeněk. *PDT 2.0 – Guide*. UFAL & CKL, 2006. Dostupné na http://ufal.mff.cuni.cz/pdt2.0/

CHIARCOS, Christian; KRASAVINA, Olga. *Annotation Guidelines, PoCoS – Potsdam Coreference Scheme*. říjen 2005. Dostupné na http://amor.cms.hu-berlin.de/~krasavio/annorichtlinien.pdf

DAVIES, Sarah; POESIO, Massimo; BRUNESEAUX, Florence; ROMARY, Laurent. *Annotating coreference in dialogues: Proposal for a scheme for MATE. Deliverable D2.1*. 1998. Dostupné na http://www.ims.uni-stuttgart.de/projekte/mate/mdag .

LEZIN, Grigory V. On automatic disclosure of referencial coherency in narrative text. In IOMDIN, Leonid L., LAUFER, Natalie I., NARINJANI, Alexandr S., SELEGEY, Vladimir P. Computational Linguistics and Intellectual Technologies. International

Conference „Dialogue 2007" Proceedings. Moskva: izdatelstvo RGGU, 2007. Dostupné na http://www.dialog-21.ru/dialog2007/materials/pdf/LezinG.pdf.

MENDOZOVÁ (MENDOZA), Imke. *Nominaldetermination im Polnischen. Die primären Ausdrucksmittel*. München. 2004. Nepublikovaná habilitační práce. Dostupné na http://www.slavistik.uni-muenchen.de/pers_pages/mendoza.htm

MENGEL, Andreas; DYBKJAER, Laila; GARRIDO, Javier M.; HEID, Uli; KLEIN, Marion; PIRRELLI, Vito; POESIO, Massimo; QUAZZA, Silvia; SCHIFFRIN, Amanda; SORIA, Claudia. MATE Dialogue Annotation Guidelines. Technical Report. 2000. Dostupné na http://www.ims.uni-stuttgart.de/projekte/mate/mdag/

NOVÁK, Václav. *Semantic Network Manual Annotation and its Evaluation.* Nepublikovaná dizertační práce. Dostupné na http://ufal.mff.cuni.cz/~novak/vn_phd_thesis.pdf

POESIO, Massimo. *The GNOME annotation scheme manual.* 2000d. Dostupné na http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm

POESIO, Massimo. Coreference. In MENGEL, Andreas; DYBKJAER, Laila; GARRIDO, Javier M.; HEID, Uli; KLEIN, Marion; PIRRELLI, Vito; POESIO, Massimo; QUAZZA, Silvia; SCHIFFRIN, Amanda; SORIA, Claudia. MATE Dialogue Annotation Guidelines. Technical Report. 2000a. Dostupné na http://www.ims.uni-stuttgart.de/projekte/mate/mdag/.

POESIO, Massimo; DELMONTE, Rodolfo; BRISTOT, Antonella; CHIRAN, Luminita; TONELLI, Sara. *The VENEX corpus of anaphora and deixis in spoken and written Italian*. Nepublikovaný rukopis. 2008. Dostupné na http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf

POESIO Massimo. *Empirical Investigation of Anaphora and Salience.* Vilem Mathesius Lectures. Prague, 2006. Dostupné na http://lectures.ms.mff.cuni.cz/video/categoryshow/index/23

RECASENS(OVÁ), Marta. *Towards Coreference Resolution for Catalan and Spanish.* Nepublikovaná diplomová práce. University of Barcelona. 2008. http://clic.ub.edu/files/dea-recasens.pdf

SEKINE, Satoshi. *Named Entity: History and Future.* 2004. Dostupné na http://www.cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf