ROZŠÍŘENÁ TEXTOVÁ KOREFERENCE A ASOCIAČNÍ ANAFORA
Koncepce anotace českých dat v Pražském závislostním korpusu

Anna Nědolužko

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Anna Nědolužko

ROZŠÍŘENÁ TEXTOVÁ KOREFERENCE A ASOCIAČNÍ ANAFORA

Koncepce anotace českých dat v Pražském závislostním korpusu

Published by Institute of Formal and Applied Linguistics
as the 10th publication in the series
Studies in Computational and Theoretical Linguistics.

Reviewer: prof. PhDr. Oldřich Uličný, DrSc.

This book has been printed with the support of the projects MSM0021620838 and LC536 of The Ministry of Education of the Czech Republic.

ISBN 978-80-904571-2-6

Obsah

Poděkování

Poděkování za vedení při přípravě práce patří především prof. PhDr. Evě Hajičové, DrSc., která mě podporovala v záměru věnovat se zvolenému tématu, byla ochotná mi vždy poskytnout konzultace, dala mi v tomto směru řadu cenných rad a také práci podrobila podrobné recenzi.

Velký dík patří také 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.

Doc. RNDr. Karlu Olivovi, Dr. vděčím za podrobnou a pečlivou recenzi, kterou napsal na dizertační práci na toto téma. Řada jeho podnětů je v této knize uplatněna.

Dále chci poděkovat Bc. Magdě Rysové a Mgr. Svatavě Škodové, Ph.D. za provedenou jazykovou korekturu, kterou vzhledem k původní jazykové kvalitě práce dané mj. skutečností, že nejsem rodilou mluvčí češtiny, považuji na nadmíru náročný úkol. Odpovědnost za všechny jazykové nepřesnosti a neobratnosti samozřejmě spočívá na mně. PhDr. Šárce Zikánové, Ph.D. děkuji za ochotu a pečlivost, se kterou pročetla a okomentovala mé přípravné verze. Vděčím rovněž svým 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 v PDT, výborné nápady při řešení problematických případů, jejich dochvilnost a zodpovědnost při odevzdávání dat.

Za vytvoření a podporu anotačního nástroje a jeho přizpůsobení potřebám anotace koreference děkuji RNDr. Jiřímu Mírovskému, Ph.D. a Mgr. Petru Pajasovi, Ph.D.

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 psaní 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.

I Úvodem

Pražský závislostní korpus 2.0 (Prague Dependency Treebank, dále PDT 2.0) je soubor velkého množství českých textů obohacených o podrobnou a mezi sebou propojenou lingvistickou informaci na morfologické, povrchově syntaktické a hloubkově syntaktické (tektogramatické) rovině.[1] Anotace na tektogramatické rovině 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 technické zprávě *Anotování koreference v Pražském závislostním korpusu* (Kučová a kol. 2003) a ve shrnujícím článku *Coreferential Relations in the Prague Dependency Treebank* (Kučová – Hajičová 2004).

Tato kniha vznikla z dizertační práce, která byla napsána a obhájena v roce 2010. Původní text dizertace byl podroben četným jazykovým opravám, obsahové změny reagují na posudky recenzentů. Dizertační práce byla dovršena v době před ukončením projektu anotace rozšířené koreference a asociační anafory na tektogramatické rovině, a veškeré číselné údaje byly měřeny na polovině dat PDT 2.0. V době vydání této práce v knižní podobě byl projekt již završen, čísla a statistiky uvedené v této knize však zůstávají stejné jako v původní verzi práce.

Při kompletním zpracování rozšířené textové koreference a asociační anafory na tektogramatické rovině PDT 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 především z oblasti teorie reference s praktickým přístupem k anotaci koreference na velkém korpusovém materiálu. Vycházíme přitom z aplikačních požadavků počítačové lingvistiky. Práce proto obsahuje poměrně rozsáhlý přehled teoretických lingvistických zdrojů k teorii reference a anaforických vztahů, které nejsou všechny využity v její praktické části. 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, Hajič a kol. 2006.

Dalším cílem 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 založeny buď na velice obecných kritériích a mají lepší mezianotátorskou shodu (projekty MUC (Hirschman 1997), ACE (Doddington a kol. 2004), MATE (Poesio a kol. 2000)), nebo mají naopak příliš specifická kritéria, která jsou špatně automaticky zpracovatelná (Passonneau 1996, rozšířená verze PoCoS (Chiarchos – Krasavina 2005) aj.). Ve zpracování koreference v 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 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).

Z hlediska prostorového umístění výrazů v textu se rozlišuje anaforické a kataforické odkazování. Při odkazování k předcházejícímu výrazu nebo výpovědi se hovoří o anaforickém odkazování. Výraz, na nějž se odkazuje, je běžně označován jako antecedent. Odkazující výraz označujeme jako koreferující člen nebo anafor.

Mezi koreferenčními výrazy nemusí být anaforické ani kataforické odkazování, i když tam často bývá.

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. Podobně používáme termíny antecedent a anafor pro elementy koreferenčního vztahu bez anaforického odkazování.

Naše práce má především praktické zaměření a je pojata jako návod anotace koreferenčních vztahů a asociační anafory na tektogramatické rovině PDT. Avšak důraz klademe také na teoretické vysvětlení konkrétních rozhodnutí, která jsme učinili v některých 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 anotovaný materiál vhodný k řešení následujících úkolů:

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

- Praktické aplikace:
  i. automatické generování 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 aplikace pro zpracování přirozeného jazyka.
- Pro evaluaci výsledků při řešení úloh uvedených výše.

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 je 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 o aplikaci probíhající anotace a jsou zde 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.

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 účastnit výrazy v oznamovacích, tázacích, rozkazovacích i negovaných větách.

V současné fázi anotace rozlišujeme původní pronominální textovou koreferenci (stručný přehled viz v III.4.1, podrobný popis viz v anotačních příručkách 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 v celém korpusu PDT[1] a týká se většiny případů pronominalizace a elips. Při anotaci pronominální koreference se vyznačovaly následující vztahy:

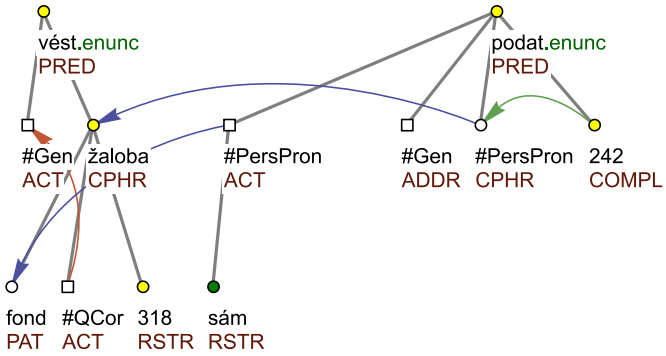
1. Textová koreference u osobních a přivlastňovacích zájmen pro 3. osobu[2] (kromě reflexivních zájmen, která byla zpracována v rámci gramatické koreference, viz. III.3). Tato zájmena mají v tektogramatickém stromě jednotné t-lemma `#PersPron`. Srov. vztah mezi *jich* a *žaloba* ve větě (1) a na obrázku 12:

   (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* ve větě (1). Při doplňování závislých valenčních doplnění všeobecným aktantem v podobě t-lematu `#Gen` nebyla

[1] K anotaci pronominální koreference viz podrobněji v anotačních příručkách Mikulová a kol. 2005, Kučová a kol. 2003.

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



Obrázek 12: Textová pronominální koreference

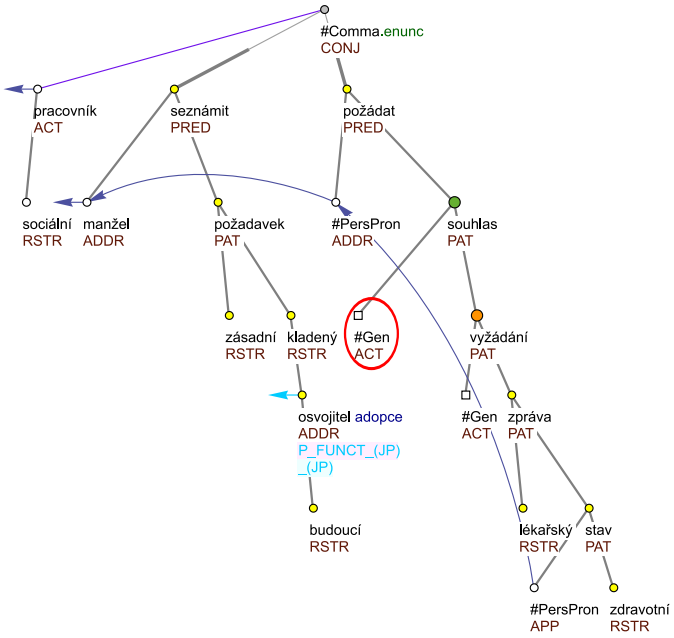
případná koreference u těchto uzlů zachycena. Srov. např. koreferenční řetězec *manžele – je – jejich* ve větě (2) a obrázek 13.

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

Při anotaci původní zájmenné koreference byla textová koreference chápána 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 případů 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.



Obrázek 13: Textová pronominální koreference

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 v PDT je rozšířením anotace pronominální textové koreference (III.4.1). Toto rozšíření se týká především typů výrazů, na které se vztahuje anotace, tj. koreferenčních párů, 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 kapitole III.4.2.3 uvádíme a rozebíráme některé problematické případy označování koreference, přičemž zvláštní pozornost věnujeme 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). Toto omezení na vzdálenost při manuální anotaci je motivováno především tím, že jinak by zřejmě došlo k velkému 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[3] (ale i v jakémkoliv jiném nástroji) více než 20 předchozích vět (textové okno zabere příliš mnoho místa, program je přetěžován informací atd.), ve výsledku tedy 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 {koreferenční vztah neanotujeme} *odmítá.*

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

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

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

(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 {koreferenční vztah neanotujeme}.
c. A také při změnách počasí {koreferenční vztah neanotujeme}, *které působí na citlivější organismus.*
d. V obdobích před a po vysvědčení {koreferenční vztah neanotujeme}.
e. V době viróz {koreferenční vztah neanotujeme}.

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 tázací 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 v 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 tyto fráze:

- Jmenné fráze v predikativní pozici kromě případů identifikačních konstrukcí, kde jmenná část přísudku může sloužit jako antecedent 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 k objektu vnějšího světa, ale sama k sobě (tzv. autonymní použití v terminologii Padučevové.[4]
  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`, koreferenční vztah neanotujeme}, *která se u nás rozmisťováním 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.*

- 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 v příkladu (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* {koreferenční vztah neanotujeme na „bod"} *v červnu.*
  b. V dubnu byla jeho hodnota rovněž 49,9 bodu {koreferenční vztah na „bod" neanotujeme}.

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 koreferovaných 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:

[4] Viz Padučevová 1985. Podobně se situace s identifikačními NP řeší v projektu PoCoS (Chiarcos – Krasavina 2005, s. 29). Tam je však řešení jednodušší, protože je prováděno na složkách, nikoliv na závislostní struktuře.

(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 odkazujících ke konkrétnímu, ale nevybranému objektu**. Srov. koreferenci NP *kolega* a *ten* v (13):[5]

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

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

3. Jinak situace s koreferencí vypadá v případě, že **výrazy odkazují generický**. 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 kapitolu II.2) atd. Reference k typu objektů se liší od reference ke konkrétním vybraným objektům tím, že se nemusí vztahovat na všechny objekty daného typu. Srov. např. ve větě (14) výraz *děti* neodkazuje ke všem reálně existujícím dětem, ale k prototypickému pojmu 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 typ objektů. Následuje-li za větou (14) věta (15), výrazy *děti* ve větách (14) a (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 ve větě (16) mluví o dětech milujících čokoládu, množina objektů, ke kterým referuje *děti*, nemusí být totožná s množinou ob-

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

jektů, ke kterým referuje *děti v* (14) – např. ne všechny děti milující čokoládu vědí, co znamená koupit, ne všechny tyto děti mají maminku atd.
V našem projektu anotace rozšířené textové koreference **anotovat textovou identickou koreferenci u generických jmenných frází však pokládáme za smysluplné,** a to z následujících důvodů:

i. Opakování stejného výrazu s generickou referencí se podílí na koherenci textu, podobně, jako je tomu v případě NP se specifickou referencí (srov. např. celý text (2) v kapitole IV.4 pojednávající obecně o dětech v dětských domovech, kde se v průběhu celého textu opakuje NP *děti* s generickou referencí);

ii. Generické jmenné fráze se mohou účastnit veškerých anaforických vztahů a podobně jako jmenné fráze se specifickou referencí mohou být pronominalizovány, (srov. paralelní syntaktické konstrukce ve větách (17) a (18)), elidovány (19) nebo opakovány s ukazovacím zájmenem (srov. např. ve větě (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 zřejmé, 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.*

iii. Hranice mezi NP se specifickou a nespecifickou referencí není vždy úplně zřetelná a v současné době není vůbec řešitelná automaticky. Pokud chceme svou 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 vztahu koreferenčního. Toto

rozhodnutí však komplikuje skutečnost, že reference k typu objektů je mnohem složitější pojem než specifická reference ke konkrétním vybraným objektům. 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) odkazovat 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) *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 mezi 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 se snažíme dodržovat extenzi koreferujících generických jmen. Vrátíme-li se k větě (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 mohou 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.

Podíváme-li se na přístupy zpracování generické koreference v projektech z oblasti počítačové lingvistiky, vidíme, že např. Lezin (2007) při realizaci projektu automatického vyhledávání referenčního 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. věty (23) a (24):

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

vs.

[5] 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.

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

Poesio (2000d) rozděluje generické a negenerické NP v atributu GENERIC definovaném pro jmenné fráze. 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 určitosti (*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*). Ke *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. věty (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, v nichž generičnost jmenné fráze nelze spolehlivě určit, u koordinačních konstrukcí, kde se generické jmenné fráze spojují s negenerickými, a u 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 odkazovat je poměrně problematickou záležitostí (viz podrobněji v III.4.2.2.1 a III.4.2.2.2). Pokud přesto uznáváme jejich schopnost odkazovat, zůstává otázkou, k čemu odkazují (srov. např. u Padučevové 1986) a jestli se mohou účastnit 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 neodkazovat, 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 kapitole III.4.2.2.1, tady chceme pouze upozornit na to, že informace o ontologickém statusu jména je pro analýzu koreferenčních párů velice důležitá.

Své rozhodnutí ohledně (ne)anotace koreference u jmenných frází s různým ontologickým statusem a typem reference shrnujeme v tabulce 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 odkazuje k anaforické NP

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

V ideálním případě potřebujeme mít pro anotaci koreference dodatečnou informaci o ontologickém statusu 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 = SPEC) a vztah mezi NP s nespecifickou (především generickou) referencí (typ = GEN)[6] (viz tabulku 17).

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

Tabulka 17: Typologie textově koreferenčních vztahů

[6] GEN – od generická reference.

Při výběru mezi textovou koreferencí typu SPEC a GEN 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 = SPEC) a generickou (typ = GEN) referencí, preferujeme „defaultní" typ SPEC.

Na začátku anotace jsme rozlišovali také typy SYN (od synonymum) pro vztah mezi koreferenčními jmennými frázemi 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 anotováno cca 10 procent PDT. Potom se však ukázalo, že vztah hyponym/hyperonym je velice nejednoznačný, do hyperonymických vztahů se dostává mnoho sporných případů, které zhoršují mezianotátorskou shodu a prodlužují dobu anotování; ve své čisté podobě se hyperonymie 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, v nichž je řídící uzel anaforu vyjádřen jiným lexémem než řídící uzel antecedentu (typ SYN), lze provést automaticky s použitím atributů tektogramatické roviny.

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

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* ve větě (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 = SPEC 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 SPEC) se mohou účastnit následující dvojice výrazů:

a) **Původní pronominální koreference**
Všechny vztahy původní pronominální koreference mají předvolený typ SPEC. V případech, kdy to neodpovídá skutečnosti, se typ vztahu následně ručně opravuje (viz III.8).
   - Antecedentem je zájmeno nebo rekonstruovaný uzel (s t-lemmaty `#PersPron`, `#Cor`, `#QCor` aj.)

- Rozšířenou koreferencí typu SPEC 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 kapitole 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 = SPEC na „Péťa“ v *a.*} *Stále na to myslel, ve škole se už nedokázal soustředit.*
[… 5 vět …]
c. Péťa {coref_text, typ = SPEC na `#PersPron` v *b.*} *skončil u Jany Drtilové.*

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

(30) *a. Informovala o tom Křesťanská misijní společnost, která* {coref_gram na „společnost“} *toto shromáždění pořádala.*
b. Společnost {coref_text, typ = SPEC 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 = SPEC na „soutěž“ v *a.*} *se spolupodílí i Profit.*

c) **Opakování stejného pojmenování s determinátorem**
Identita referentů je vyjádřena pomocí textových identifikátorů. Srov. příklad, kdy se se stejnou NP v anaforické pozici používá ukazovací zájmeno:

(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 = SPEC na „žena“} *s dětmi dělat?*

d) **Opakování různých podstromů při stejném řídícím uzlu**
Jako coref_text, typ = SPEC 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 = SPEC 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 synonymní:

(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 = SPEC 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 odkazují 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 = SPEC 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 = SPEC 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 = SPEC 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 SPEC (coref_text, typ = SPEC). Srov. větu (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 = SPEC 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 SPEC 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 odkazuje 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 sice antecedent má nespecifickou referenci, ale pak se s ním operuje jako s konkrétním vybraným objektem, čili se vytváří fiktivní svět daného diskurzu, ve kterém se daný objekt chová jako existující a reálný. Srov. např. koreferenci výrazů *podnik* a *úřad* ve větě (39) a *velice pozorný člověk* ve větě (40):

(39) *Například muž, který pracuje v nějakém velkém podniku, se zakouká do sekretářky ve stejném podniku* {coref_text, typ = SPEC „podnik“} *a začnou se scházet v nějaké kavárničcce stranou od toho úřadu* {coref_text, typ = SPEC „podnik“}.

(40) *Přesto si značky mohl všimnout jen někdo velice pozorný* […] *a ani ten velice pozorný člověk* {coref_text, typ = SPEC „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 = SPEC 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 = GEN 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 = GEN)

O jmenné skupině se říká, že je použita genericky, jestliže jejím referentem je klasický, vzorový, prototypický představitel dané třídy (např. příklady (43)–(45)), odkaz na libovolný element dané třídy (např. (46)) nebo na reprezentativní 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 použito v konstrukci „*X* jako <typické> *Y*", „*X* jako druh (forma) *Y*". Například ve větě (48) má *vlak* generickou referenci, protože větu (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 rádi zmrzlinu.*

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

Pravidlo o anotaci textové koreference u generických NP:

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

Textovou koreferenci typu GEN zaznamenáváme:

1. U generických jmenných frází v singuláru a v plurálu
   a) **Vyjádřených stejným pojmenováním**, srov. NP *českým exportérům* ve větě (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 = GEN na „exportér" v (53)a} *se dotýkají zejména oblasti obchodu s potravinami a zemědělskými produkty.*

   b) Pokud je anaforický člen **pronominalizace nebo aktuální elipsa generického antecedentu**. V tomto případě měníme automaticky předvolený typ SPEC na typ GEN.[7] Srov. aktuální elipsu NP *měřicí přístroje* ve větě (54) a pronominalizaci NP *droga* ve větě (55):

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

      (55) *Droga je tedy tak účinná, že ten, kdo ji* {coref_text, typ = GEN 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.*

[7] 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.)

b. Doufejme, že linka si časem vydobude mezi dětmi {coref_text, typ = GEN 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 = GEN na „DPH“} *se letos stanu plátcem daně.*

e) Pokud **anaforická generická NP je hyperonymum ve vztahu k antecedentu**. V tom případě platí pravidlo, že použití aktualizátoru pro zachování koreference s antecedentem je nezbytné. Srov. např. (58):[8]

(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 = GEN na „kůrovec“}, *začínal rojit.*

f) **U uzlů závislých na „kontejnerech“,**[9] 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* ve větě (59)a–b a na obrázku 14.

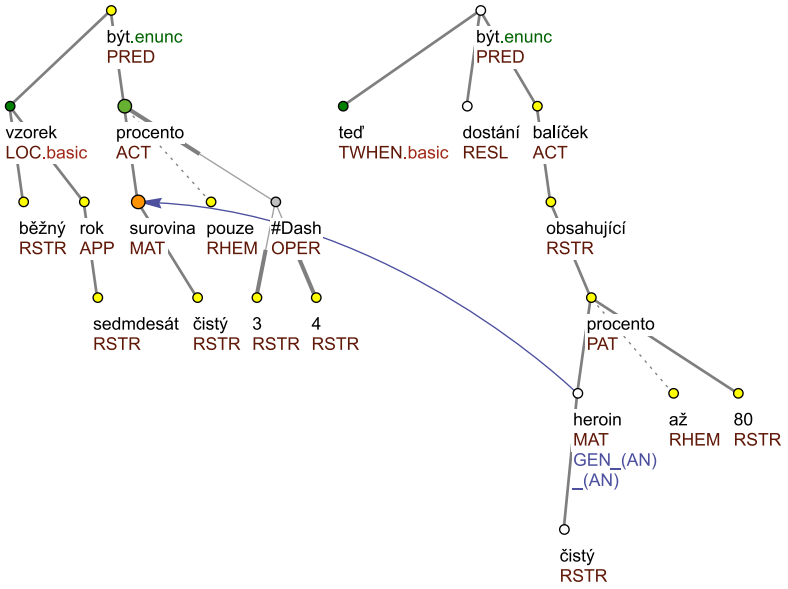
(59) *a. V běžném vzorku sedmdesátých let byla pouze 3–4 procenta čisté suroviny.*
b. Nyní jsou k dostání balíčky obsahující až 80 procent čistého heroinu {coref_text, typ = GEN na „surovina“}.

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 = GEN na „křesťan“ v (60)a} *z různých sborů a církví českých zemí a delegace křesťanů* {coref_text, typ = GEN 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.*

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

[8] Příklad ze SYN2000.
[9] K vysvětlení pojmu „kontejner“ v anotaci tektogramatické roviny viz Mikulová a kol., 2005, s. 809.



Obrázek 14: Koreference generických výrazů závislých na kontejnerech

(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 = GEN 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 = GEN na „rodina“ v a.} *dozvěděla, že se nám ozval.* (VL)
c. Linka by neměla rodinu {koreferenční vztah neanotujeme} *nahrazovat, ale doplňovat.*

Pokud však věta (62)a pokračuje větou (62)c, koreferenční vztah mezi *rodina* ve větě (62)c a *rodina* ve větě (62)a neanotujeme. Ve větě (62)a má *rodina* nereferenční negenerickou interpretaci, zatímco ve větě (62)c je *rodina* použita genericky.

Ani při poměrně širokém chápání pojmu koreference tyto dvě NP nejsou koreferenční.
Je zřejmé, že vztah mezi negenerickými nereferenčními NP nepřispívá tolik ke koherenci textu a možná by bylo logičtější takové vztahy vůbec nezaznamenávat – koreference v úzkém smyslu zde není (obě NP odkazují k nevybranému objektu), o anaforický vztah také nejde (jinak by se tento vztah označoval jako koreference mezi NP se specifickou referencí, viz III.4.2.1.1). Avšak hranice mezi nereferenčními negenerickými 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 typu GEN.

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

V textu se koreferenční řetězce typu SPEC a typu GEN mohou prolínat, tj. některé hrany jednoho řetězce mohou být označeny GEN, jiné však SPEC, 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 = SPEC na „kluk“ v *a.*} *nikdo nevede, nehlídá ani nevychovává.*
c. Kdo to {specifická reference, coref_text, typ = SPEC na „chlapec“ v *b*} *vlastně je?*
d. Německé chlapce {nespecifická reference, coref_text, typ = GEN na „ten“ v *c.*} *jsme již nezastihli.*
e. Duchcov byl posledním místem, odkud `#PersPron` {specifická reference, coref_text, typ = SPEC na „chlapec“ v *d.*} *byli těsně před naším příjezdem odvezeni zpět do Německa.*

V následujícím případě (64)a–d má NP *chicle* 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 = GEN na poslední „chicle“ v *a.*} *a Adams technické schopnosti.*
c. Asi rok se Adams a jeho nejstarší syn snažili – chicle {coref_text, typ = GEN 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 = SPEC na „chicle“ v *c.*} *ještě zbylo, hodí do řeky.*

Srov. také koreferenci v páru *Romové – tento národ*, kde má první výraz generickou referenci a druhý specifickou:[10]

(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 = GEN 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 = SPEC 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 tak jednoduché není.

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 označuje reálné předměty vnějšího světa, denotáty, 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. o*blečení* (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 zastoupit své 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.

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

Summary

The purpose of this book is to describe the 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 SPEC) and relations between NPs with generic reference (coref_text, type GEN). 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 Uni-

versity 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 SPEC), 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, látka, 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, která jsou použita 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 „-“ nebo „–“;
`#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

Seznam grafů

Literatura

ADAMEC, Přemysl. Funkcii ukazateľnych mestoimenij v češskom jazyke v sravnenii s russkim. In ŠIROKOVÁ, Alexandra G.; GRABJE, Vladimír. *Sopostaviteľnoje izučenije grammatiki i leksiki russkogo jazyka s češskim jazykom i drugimi slavjanskimi jazykami*. Moskva: izdateľstvo MGU, 1983, s. 173–190.

ADAMEC, Přemysl. K prostředkům textové syntaxe v současné češtině. In *Přednášky z XXX. Běhu LŠSS, 1986*; UK, Praha, 1988, s. 105–115.

ADAMEC, Přemysl. K vyjadřování referenční určenosti v češtině a ruštině. *Slovo a slovesnost*, 1980, roč. 41, s. 257–264.

ADAMEC, Přemysl. Različija v vyraženii anaforičeskich otnošenij meždu russkim i češskim jazykami. *Russkij jazyk za rubežom.* Moskva, 1984. s. 73–78.

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.

ARUŤUNOVOVÁ (ARUŤUNOVA), 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.

BEĽSKIJ, 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, Rosanna. K voprosu ob opredelennom artikle v slavjanskich jazykach: reźjanskij govor. In DULIČENKO, Alexandr D. (ed.). *Jazyki malye i bolšie.* 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.

BOGUSLAVSKÁ, Olga Ju.; MURAVJEVA, Irina A. Mechanizmy anaforičeskoj nominacii. In KIBRIK Alexandr E.; NARIŇJANI Alexandr S. (eds.). *Modelirovanie jazykovoj dejateľnosti v intellektuaľnych sistemach*. Moskva: Nauka, 1987, s. 78–128.

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.

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, 1977.

COHEN, Jacob. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, roč. 20(1), 1960, s. 37–46.

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.

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.

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. 1985.

FREGE, Gottlob. Über Begriff und Gegenstand. *Vierteljahresschrift für wissenschaftliche Philosophie*. Leipzig, roč. 16, 1892, s. 192–205.

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 funkcionaľnoj grammatiki IV. Subjektnosť. Objektnosť. Kommunikativnaja perspektiva vyskazyvanija. Opredelennosť/neopredelennosť*. 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). S*emantik. 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 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.

HLAVSA, Zdenek. K protikladu určenosti v češtině. *Slovo a slovesnost*, roč. 33, 1972, s. 199–203.

HRBÁČEK Josef . *Nárys textové syntaxe spisovné češtiny*. Praha: Trizonia, 1994.

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.

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.

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.

KRIPKE, Saul. *Naming and Necessity*. Cambridge, Mass.: Harvard University Press, 1980.

KOMÁREK, Miroslav. Sémantická struktura deiktických slov v češtině. *Slovo a slovesnost*, roč.39, 1978, s. 5–14.

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 otglagoľnych imen *Naučno-techničeskaja informacija*, seria 2, roč. 12, 1981, s. 17–22.

KŘÍŽKOVÁ, Helena. Zájmena typu *ten* a *takový* v současných slovanských jazycích. *Slavica Slovaca*, roč. 6, 1971, č.1, s. 15–30.

KUČOVÁ, Lucie; HAJIČOVÁ, Eva. Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of 5th Discourse Anaphora and Anaphor Resolution Colloquium*. Edicoes Colibri, 2004.

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.

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.

MEĽČUK, Igor. A. *Opyt teorii lingvističeskich modeley „Smysl ↔ Text"*. Moskva: Nauka, 1974 (2. vydání 1999).

MIKULOVÁ, Marie; BÉMOVÁ, Allevtina; HAJIČ, Jan; HAJIČOVÁ, Eva; HAVELKA, Jiří; KOLÁŘOVÁ, Veronika; KUČOVÁ, Lucie; LOPATKOVÁ, Markéta; PAJAS, Petr; PANEVOVÁ, Jarmila; RAZÍMOVÁ, Magda; SGALL, Petr; ŠTĚPÁNEK, Jan; UREŠOVÁ, Zdeňka; VESELÁ, Kateřina; ŽABOKRTSKÝ, Zdeněk. *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.

MLADOVÁ, Lucie. *Diskurzní vztahy v češtině a jejich zachycení v anotovaném korpusu*. Nepublikovaná diplomová práce. Praha: Filozofická fakulta Univerzity Karlovy, 2008.

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.

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.

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.

NEDOLUZHKO, Anna. Razmetka koreferencii na sintaksičeski annotorovannom korpuse češskich tekstov. In KIBRIK, Alexandr E. a kol. (eds.). *Computational Lingu-*

istics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2009". Issue 8 (15). 2009, Moskva: RGGU, s. 332 – 339.

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* 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. 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.

NIKOLEJEVOVÁ, Tatiana M. Kategorija opredelennosti-neopredelennosti v slavjanskich i balkánských jazykach. Moskva: Nauka, 1979.

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.

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.

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. 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 neékstensionaľnych kontekstach. *Naučno-techničeskaja informacija*, ser. 2, roč. 1, 1983, s. 24–29.

PADUČEVOVÁ (PADUČEVA), Elena V. Nacionaľnyj 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. O referencii jazykovych vyraženij s nepredmetnym značenijem. *Naučno-techničeskaja informacija*, ser. 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 koreferentnosť. In: USPENSKIJ, Boris (ed.). *Voprosy kibernetiki. Problemy razrabotki formaľnoj modeli jazyka*. Moskva: Nauka, 1988, s. 71–88.

PADUČEVOVÁ (PADUČEVA), Elena V. *Vyskazyvanije i jego sootnesennosť s dejstviteľnostju*. Moskva: Nauka, 1985.

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. 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. 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. The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL.* Boston, duben, 2004a.

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; 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.

POESIO, Massimo; BRUNESEAUX, Florence; ROMARY, Laurent. The MATE metascheme for coreference in dialogues in multiple language. In *Proceedings of the ACL Workshop on Standards for Discourse Tagging*. Maryland, červen, 1999.

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; MEHTA, Rahul; MAROUDAS, Axel; HITZEMAN, Janet. Learning to resolve bridging references. In *Proceedings of ACL.* Barcelona, červenec, 2004b.

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; 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; VIEIRA, Renata. A Corpus-based Investigation of Definite Description Use. *Computational Linguistics*, roč. 24, č. 2, 1998, s. 183–216.

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. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. In *Language Resources and Evaluation*. 2010.

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.

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.

SEARLE, John R. *Speech acts: An essay in the philosophy of language*. UK: Cambridge university press. 1969.

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. 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.

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.

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.

STEDE, Manfred. The Potsdam Commentary Corpus. In *Proceedings of the ACL – 2004. Workshop on Discourse Annotation.* Barcelona, 2004, s. 96–102.

STEPANOV, Jury S. *Imena, predikaty, predloženija (semiologičeskaja grammatika)*. Moskva: Editorial URSS, 2004.

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.

ŠMELEV, Alexey D. *Opredelennosť – neopredelennosť v nazvanijach lic v russkom jazyke*. nepublikovaná dizertační práce. Moskva, 1984.

ŠMELEV, Alexey D. *Referenciaľnye mechanizmy russkogo jazyka*. Tampere: Slavica Tamperensia, 1996.

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. *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; POESIO, Massimo. An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics*, roč. 26, č. 4, 2000, s. 539–593.

VIEIRA, Renata; TEUFEL, Simone. Towards resolution of bridging descriptions. In *Proceedings to 35th Annual Meeting of the Association for Computational Linguistics.* Saarbrücken, Germany, 1997.

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

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 .

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

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. 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. *Empirical Investigation of Anaphora and Salience*. Vilem Mathesius Lectures. Prague, 2006. Dostupné na http://lectures.ms.mff.cuni.cz/video/categoryshow/index/23

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. *The GNOME annotation scheme manual*. 2000d. Dostupné na http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm

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

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