

M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A
P R A H A

**ZPRÁVA K ANOTOVÁNÍ ROZŠÍŘENÉ TEXTOVÉ KOREFERENCE A
BRIDGING VZTAHŮ V PRAŽSKÉM ZÁVISLOSTNÍM KORPUSU.**

ANJA NĚDOLUŽKO
12. ZÁŘÍ 2007

U N I V E R Z I T A K A R L O V A V P R A Z E

Obsah

ÚVOD	5
A. GRAMATICKÁ KOREFERENCE	5
B. TEXTOVÁ KOREFERENCE	6
B.1. Dřívější zájmenná koreference – pronominalizace a elipsa	6
B.2. Rozšířená anotace textové koreference	7
B.2.1. Slovnědruhovú charakteristika koreferovaných párů	7
B.2.2. Typologie textově koreferenčních vztahů	7
B.2.2.1. Koreferenční vztah mezi NP se specifickou referencí, kde koreferující člen není synonymum ani hyperonymum antecedenta.	8
Koreference osobních zájmen v dialogických textech.	8
Koreference otázkového slova a odpovědi v dialogických textech.	9
B.2.2.2. Koreference synonymních NP (SYN).	9
B.2.2.3. Koreference hyponymu a hyperonymu (ER).	10
??? K otázce výběru antecedentů u řetězců se specifickou referencí:	11
B.2.2.4. Koreference nereferenčních a generických NP (NR).	12
??? Hraniční případy u koreference NP se specifickou referencí a NP s textovou koreferencí typu NR.	12
??? Hraniční případy koreference s typem NR a něčím, co nemusí být jako koreference označeno	13
C. BRIDGING VZTAHY	15
C.1. bridging-vztah část-celek (PART)	16
Sporný příklad 1 (odkaz k něvyjádřenému členu):	17
Sporný příklad 2 (velká vzdálenost mezi členy (28 vět) a velice sporná sémantická souvislost – může být jenom náhoda):	17
C.2. bridging-vztah množina-podmnožina/element množiny (SET)	17
Sporný příklad 1 (vztah SET uvnitř jedné věty).	18
Sporný příklad 2 (otázka nutnosti a hloubky interpretace).	18
Sporný příklad 3 (označování některých bridging vztahů u nereferenčních NP)	19
Odkazování k více uzlům	19
??? Hraniční případy mezi bridging typy SET a PART.	20
C.3. bridging-vztah sémantického protikladu (CONTRAST)	20
C.4. bridging-vztah zatím neterminovaný (REST)	21
Ad hypertextematické propojení textu.	22
D. SPECIÁLNÍ TYPY REFERENCE (COREF_SPECIAL)	22

Hraniční případy mezi typem coref_special, typ segm a bridging anaforou, typ SET.	23
Ad named-entities.	23
PROBLÉMY A HODNOCENÍ PŘEDPOKLÁDANÉ KVALITY BUDOUCÍ ANOTACE	24
ROVINA ANALÝZY DISKURZU	24
Anotováno dat:	25
Statistika počtu vztahů:	25
Celkem	25
Srovnání s předchozí verzí od Lucky:	25
SEZNAMY TYPŮ S PŘÍKLADY:	26

Úvod

Existující anotace koreference v PDT 2.0. vychází z pojmu reference jazykových jednotek a dělí se na gramatickou a textovou koreferenci. Oblasti gramatické koreference a pronominální textové koreference jsou kompletně zpracovány na tektogramatické rovině a popsány ve „velkém manuálu“.

Další krok je rozšíření textové koreference o jiné vztahy než pronominalizace a zavádění třetí koreferenční oblasti – bridging vztahů. Tyto kroky se pokusím v následující zprávě stručně představit.

Pozor! Jsme si vědomi toho, že termín koreference implikuje pouze identitu referentů objektů, přesto pro zjednodušení občas užíváme termínu koreference i pro případy bridging-vztahů a mimotextového odkazování.

A. Gramatická koreference

Koreferovaný člen je určován na základě gramatických pravidel daného jazyka. Všechny šípky, co tam byly, zůstaly. Nic jsem nepřidala ani nesmazala. Ani jsem je nekontrolovala.

Problém anotace rozšířené koreference, který je spojen s již oannotovanou gramatickou koreferencí je však v tom, kam povedou nové šípky – (a) na původní lexikálně vyjádřený uzel, nebo (b) na uzel, od kterého vede šípka s gramatickou koreferencí na tento lexikálně vyjádřený původní uzel. Příklady:

(13/In94204_107) Linka bezpečí 855 44 33 , kterou od zítřejšího dne provozuje v Praze nadace Naše dítě , byla zřízena především s úmyslem pomoci fyzicky i duševně týraným a pohlavně zneužívaným dětem .[...]

(17/In94204_107) Na toto telefonní číslo {coref_text,typ_0 na „který“} však mohou samozřejmě zavolat všichni kluci a děvčata , kteří se ocitnou ve svízelné situaci.

Při anotaci jsem vztahovala člen s identickou koreferencí na poslední koreferencí připojený uzel, tj. v daném případě na kterou, nikoliv na Linka bezpečí 855 44 33, aby se dodržel koreferenční řetězec. Pokud jde o bridging vztah, tam bych propojovala mezi sebou jenom lexikálně vyjádřené jednotky. Srov. např.:

(17/In94204_107) Na toto telefonní číslo však mohou samozřejmě zavolat všichni kluci a děvčata , kterí se ocitnou ve svízelné situaci.

(18/In94204_107) Ptali jsme se několika {bridging, typ_SET na „kluci a děvčata“}, jestli by takového kamaráda po telefonu považovali za dobrou věc .

Podobné problémy nastávají s odkazy na nevyjádřené rekonstruované #PersPron, kde už nejde o gramatickou koreferenci ale o textovou. Tam to řeším podobným způsobem. (Viz dále u textové koreference)

B. Textová koreference

Na dnešní etapě anotace rozlišujeme dva typy – „dřívější“ pronominální anaforu (kataforu) a pozdější rozšířenou, ještě ne úplně ustálenou a provedenou jenom na vybraných souborech. Základní princip textové koreference, jak původní tak i rozšířený, je identita referentů koreferovaného a koreferujícího členu.

Jako koreferenční neoznačujeme:

- a) Vztah mezi jednotlivými členy apozice (z podstaty apozice vyplývá, že jednotlivá pojmenování jsou koreferenční).
- b) Vztah mezi subjektem a predikátovou částí výpovědi Je to zřejmé zejména u přísudku jmenného se sponou (*Petr je lékař – Petr ... lékař*). Predikace je usouvztažnění jednoho slova k druhému, skutečnosti ke skutečnosti, takže by to bylo zbytečné.

B.1. Dřívější zájmenná koreference – pronominalizace a elipsa

(dědictví PDT 2.0, v datech už je zaznačená jakožto textová koreference. Jde o osobní zájmena a ukazovací zájmeno „ten“, které není v atributivní pozici, a také o aktuální elipsu, která je v tektogramatickém stromě označena rekonstruovaným #PersPron) zůstala z obsahové stránky beze změn (moc jsem to ani nekontrolovala), až na jednu důležitou technickou drobnost. V rozšířené anotaci koreference jsme zavedli klasifikaci identické koreference na několik typů (viz dále v textu). Je to jistý technický zákrok do anotační struktury, tedy každá tmavomodrá šípka teď má automaticky označen jeden ze čtyř typů. Pokud nic není přidáno ručně, má uzel předvolený typ nula (koreferenční vztah mezi NP se specifickou referencí, kde je druhý člen zájmeno nebo podstatné jméno). Z toho plyne, že všechny dříve označené textové koreference mají teď označen tento typ nula, i když občas poukazují k nereferenčním NP. Srov např.

(54/In94204_107) *Dítě* je ještě z formy , ale už #PersPron musí plnit doma i ve škole spoustu úkolů a působí *mu* to problémy .

Může se zdát, že v daném případě označovat vztah mezi NP *dítě* a NP #PersPron a *mu* jako vztah mezi nereferenčními NP je zbytečné, protože funguje jako normální anaforický vztah, avšak NP *děti* z daného příkladu s nereferenčním statutem prochází celým textem a chová se jinak, než by se chovala, kdyby měla specifickou referenci (například se ani jednou nevyskytuje s odkazovacím zájmenem apod.)

Při anotaci jsem ten problém zatím systematicky neřešila (někde jsem to změnila, kde mi to výloženě vadilo, někde jsem si toho vůbec nevšimla), ale bude nutno to udělat při další anotaci většího počtu souborů. Pokud to neuděláme, zkreslíme statistiku koreferenčních vztahů NP se specifickou a generickou referencí. Pokud vůbec zachováme typ NR, dá se to udělat buď ručně při anotaci rozšířené koreference, nebo automaticky tak, že pokud #PersPron vede na uzel, který je spojen s předchzím uzlem textovou koreferencí a má typ NR, pak i vztah #PersPron k jeho uzlu má typ NR.

B.2. Rozšířená anotace textové koreference

– jiný vztah než pronominalizace a elipsa.

B.2.1. Slovnědruhová charakteristika koreferovaných párů

Na etapě rozšířené anotace textové koreference označujeme koreferenční vztah i mezi jinými páry, než substantivum-zájmeno. Jsou následující možnosti:

- i. koreferující člen je vyjádřen **substantivem**:
substantivum¹ – substantivum (*snímek – fotografie, Petr – ten kluk, pí Novotná – mluvčí* apod.)
zájmeno – substantivum (*on – Kolář*)
elidovaný #PersPron – substantivum (#PersPron - *Kolář*)
příslovce – substantivum (*tady – v Praze*)
sloveso – substantivum (*prodávat – prodej, plavat – tento proces*)
- ii. koreferující člen je vyjádřen **adverbiem**:
antecedent (A)² – příslovce (*Praha – (#PersPron) – (ta) – (odtud) – tam*)
- iii. jeden z členů páru je **adjektivum**. Tuto skupinu беру zvlášť, protože u adjektiv (aspoň zatím, ale možná i vůbec) koreferenci důsledně neanotujeme. Především proto, že se u adjektiv vůbec těžko jakákoliv reference určuje – většinou nereferují ale predikují. V některých případech se však bez takové anotace těžko obejdeme. Můžeme vyčlenit 2 menší skupiny, u kterých bych nabízela koreferenci anotovat:
 - a) adjektivum je přivlastňovací (*maminka – maminčin názor*) – pak je to jako přivlastňovací zájmeno, kde gramatickou koreferenci normálně anotujeme (*maminka – svůj názor*). Adjektivum může mít koreferenční šípku, i když přivlastňovací význam je vyjádřen u gramaticky neutrálního adjektiva (*palácový – palác, dětský* s významem „dítěte“ např. v „*dětská mysl*“ v následujícím příkladě – neumím to pořádně terminologicky pojmenovat) (21/In94204_107) Co se může dospělému zdát zanedbatelnou záležitostí, naroste v dětské {bridging, typ_CONTRAST, na „dospělý“} myslí třeba i do tragických rozměrů .
 - b) NP se skládá z adjektiva a substantiva, přičemž lexikální sémantika NP je vyjádřena adjektivem a subst. nese jenom formální a gramatickou funkci (*ekonomie – ekonomických nauk* nebo v jiném pořadí)

Sloveso v roli koreferujícího člena nevystupuje.

B.2.2. Typologie textově koreferenčních vztahů

Na dané etapě rozlišujeme 4 druhy vztahů mezi koreferovaným a koreferujícím členy v páru NP spojených textovou koreferencí:

- | | |
|-------------------------------------|--|
| 0 | – vztah mezi koreferenčními NP se specifickou referencí, kde koreferující člen není ani synonymum ani hyperonymum antecedenta (viz B.2.2.1.) |
| SYN (od <i>synonymum</i>) | – vztah mezi synonymními NP se specifickou referencí (viz B.2.2.2.) |
| ER (od <i>hyperonymum</i>) | – vztah mezi NP se specifickou referencí, kde druhý člen je lexikální hyperonym ve vztahu k prvnímu (viz B.2.2.3.) |
| NR (od <i>nereferenční</i>) | – vztah mezi nereferenčními a generickými NP (viz B.2.2.4.) |

¹ antecedent = substantivum, zájmeno apod. zde a dále vždy znamená, že tento slovní druh je kořen příp. podstromu antecedenta.

² Antecedent ∈ {substantivum, zájmeno, elidovaný člen, příslovce}

B.2.2.1. Koreferenční vztah mezi NP se specifickou referencí, kde koreferující člen není synonymum ani hyperonymum antecedenta.

Tento vztah je základní a v anotaci je předvolený. Oba jeho členy mají specifickou referenci (odkazují ke konkrétnímu existujícímu, reálnému referentu a objektu skutečnosti.) Srov např.

(6/In9413_006) Jeho dojetí znásobila při vyhlašování přítomnost [...] pořadatelů soutěže - českého manažerského centra v Čelákovících .

(7/In9413_006) Na letošním ročníku soutěže se spolupodílí i Profit .

(29/In9413_006) Začal jsem provozováním hospody, která {coref_gram, na „hospody“} byla mnohokrát vykradena. [...]

(32/In9413_006) Hospoda {coref_text, typ_0 na „který“} byla jen startem, polem k podnikání s masem a masnými výrobky.

Mezi dvěma uvedenými příklady je jistý rozdíl v „konkrétnosti“ reference podtržených NP. Ve větě (32) jsou možné v podstatě dvě interpretace NP hospoda – jako NP se specifickou referencí (ta konkrétní hospoda, kterou podnikatel provozoval) a jako NP s generickým referenčním statutem (hospoda jako taková – podnikatel chtěl pořídit něco jako hospodu, aby poznal svět podnikání.) Takových nejednoznačných případů je nečekaně moc. Neumím to zatím elegantně řešit. Pokud však není úplně zřetelná nereferenční interpretace, anotovala bych to jako NP se specifickou referencí.

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

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 s dětmi dělat? (Fuks, L., Spalovač mrtvol)

Velká výhoda rozšířené anotace identické koreference je v tom, že můžeme propojit nejenom pár NP - #PersPron, ale i opačný směr #PersPron – NP, pokud se v koreferenčním řetězci vystupuje. Srov. posloupnost vět (24)-(34)/In94204_107:

(24) Sedmiletý Pěťa se půl roku neuvěřitelně trápil, že má AIDS. [...]

(29) #PersPron Stále na to myslel, ve škole se už nedokázal soustředit. [...]

(34) Pěťa {coref_text, typ_0 na #PersPron} skončil u Jany Drtilové.

Koreference osobních zájmen v dialogických textech.

V PDT 2.0 se koreference neoznačuje u osobních zájmen 1. a 2. osoby. Propojování osob mluvčích v dialogickém textu se tedy neprovádí. Srov. např.

(1/In9413_006) #PersPron.ACT Začal podnikat a vystřízlivěl [...]

(4/In9413_006) #PersPron.ACT (žádný koreferenční odkaz) Byl jsem úplně naměkko, neschopen mluvit.

(5/In9413_006) Tak hodnotí Petr Chodura, podnikatel (žádný koreferenční odkaz) z Ostravy, první momenty po oznámení, že se stal Vynikajícím podnikatelem roku 1993. [...]

(20/In9413_006) Takže #PersPron.ACT (žádný koreferenční odkaz) jste se cítil schopen" jít do toho"?

V real-time dialozích, které se anotují na anglickém materiálu (Cinková a kol.) je přítomen exoforický odkaz na entity-ID. Tady je však dialog zahrnut do textu. Nemůžeme propojit třetí osobu z úvodního textu s první osobou přímé řeči, především proto, že mezi přímou řečí a řečí autora nemohou fungovat anaforická pravidla. Ale mezi replikami v dialogu anaforická pravidla fungovat mohou: (*Seznámil jsem se včera s hezkou holkou. – No, a co ti ta holka říkala?*) Avšak v rozšířené anotaci koreference na tektogramatické rovině nebudeme propojovat mluvčích v repliky přímé řeči. Necháme to na příp. budoucí rovinu kognitivní (diskurzivní), která se připravuje(?).

Koreference otázkového slova a odpovědi v dialogických textech.

Jde o dialogy typu:

(45/In94204_107) Kdy děti nejvíce volají [...]
(49/In94204_107) Podle zkušeností ze zahraničí se dá předpokládat, že největší frekvence telefonátů nastane vždy mezi 16 . až 18 . hodinou.
(50/In94204_107) A také při změnách počasí , které působí na citlivější organismus .
(51/In94204_107) V obdobích před a po vysvědčení .
(52/In94204_107) V době viróz .

Pro koherenci textu vztah mezi otázkovým slovem a částí odpovědi, která na tu otázku reaguje, zachycení vztahu mezi nimi je velice důležité. Je to však otázka přesahující tektogramatickou rovinu a patří spíše do roviny diskurzu. Při rozšířené anotaci tektogramatické roviny tento vztah tedy nezachycujeme, necháváme ho stejně jako koreferenci osobních zájmen 1. a 2. osoby na příp. budoucí rovinu kognitivní (diskurzivní), která se připravuje(?).

Pozor! V dialogickém textu však běžně označujeme jiné vztahy, než koreference osobních zájmen 1. a 2. osoby a otázkové slovo – odpověď, jde-li o identickou koreferenci nebo bridging. Příklady jsou uvedeny v odpovídajících kapitolách bez zvláštního odkazu na to, že je to dialogický text.

B.2.2.2. Koreference synonymních NP (SYN).

Koreferenční vztah je vyjádřen synonymním párem. Vyjadřujeme jenom u párů se specifickou referencí, abychom předešli kombinaci příznaků a z důvodů jiného chování synonymních koreferujících členů (viz vysvětlení v B.2.2.4.). Další odůvodnění a podrobnější vysvětlení podám, až na to téma nasbírám dostatek názorných příkladů – nebo pokud to nevyjde, tak to třeba zrušíme.

(13/In9413_006) Po vojně začal v Masokombinátu v Ostravě - Martinově. [...]
(21/In9413_006) Ve státním podniku {coref_text, typ_SYN na „masokombinát“} mne ubíjel stereotyp a nepružnost.

Srov. také příklady z korpusu SYN2005 (jen jako ukázka):

Z koupelny tiše bzučela automatická pračka a z rádía Andreu značně znervózňoval art-rock , ale neodvážila se přístroj

přeladit , protože byl tak složitý , a taky věděla , jak moc si na té stereosoustavě její muž zakládá . (Zapletal, Z., Půlnoční běžci)

Přijít o všechno kvůli jakési nezletilé kurvičce ! Nebylo moudřejší sem tam se s ní vyspat a šmytec ? Určitě trpce litoval , že coby k té rajdě si nechal přerůst přes hlavu tak , že přišel o plody celé poloviny svého poctivě pracovitěho života . (Frýbová, Z., Hrůzy lásky a nenávisti)

Chlap je z Prahy , klidně může zasedat v koordinačním centru nebo být poradcem bůhví koho , takže pozor , tím vtípkováním si tě taky může prověřovat . . . Skřípavě se zasmál a řekl : " A taky , chválabohu , hned tak nepochováme . Ten hoch má tuhý kořínek , ten má sílu , ten má elán . . . (Frýbová, Z., Hrůzy lásky a nenávisti)

B.2.2.3. Koreference hyponymu a hyperonymu (ER).

Koreferující člen je vyjádřen hyperonymním výrazem ve vztahu k antecedentu. Pojmy hyperonym a hyponym chápeme v širokém smyslu se zaměřením k referenci.

Tento vztah vyjadřujeme jenom u párů se specifickou referencí, abychom předešli kombinaci příznaků a z důvodů jiného chování synonymních koreferujících členů (viz vysvětlení v B.2.2.4.). Další odůvodnění a podrobnější vysvětlení podám, až na to téma nasbírám dostatek názorných příkladů – nebo pokud to nevyjde, tak to třeba zrušíme. Předpokládám ve většině případů obligatorní výskyt u anaforická NP ukazovacího zájmena.

Srov. pokračování koreferenčního řetězce (24)-(34)/In94204_107, uvedeného v B.2.2.1.:

(34/In94204_107) Péťa skončil u Jany Drtilové.

(35/In94204_107) Všechno nakonec dobře dopadlo , ale tohle dítě {coref_text, typ_ER na „Péťa“} zbytečně prožilo půl roku strachu a děsivých představ .

Zobecňující substantiva se vyskytují také jako koreferující člen v případě, když antecedent je celá věta. Srov.:

(21/In94204_107) Co se může dospělému zdát zanedbatelnou záležitostí , naroste v dětské mysli třeba i do tragických rozměrů .

(22/In94204_107) S takovými případy {coref_text, typ_ER na kořen stromu „narůst“} se během své třicetileté praxe setkala mnohokrát.

Srov. také příklady z korpusu SYN2005 (jako ukázka, jak jsem na to přišla a obligatornosti zájmena):

Protože tenhle Adolf Hitler nebyl vůdce velkoněmecké říše, ale pták druhu tučňák královský. A to jméno dostal vlastně dodatečně. (Zábrana, J., Vražda v zastoupení)

[...] a kolem trůnu duha jako smaragdová . Mari odstrčí židli , na které sedí, a postaví se." Duha?" Kněz přiloží prst k

tomu slovu, aby nezapomněl, kde skončil. (Ludva, R., Jezdci pod slunečníkem)

A co to je?" "Slon, Kájo!" "Tady u nás neběhá, že ne? Tatínek takové zvíře ještě nikdy domů nepřinesl." "Tady nežije , Kájo . To zvíře žije v cizích krajinách . (Háj, F., Školák Kája Mařík)

Často také s časovými údaji:

A tak jsme ten rok vyjeli o něco dříve - koncem června . V tu dobu tam je sice ještě moc velká zima na koupání, ale ryby berou výborně . (Haňka, L., S puškou a udicí Severní Amerikou)

* * *

??? S principiální otázkou označovat nebo neoznačovat koreferencí v daném konkrétním páru NP se setkáváme i u jmen s jasnou specifickou referencí. V případě, když v textu nejsou žádné další prostředky koheze a oba členy páru (rozhodující je ten druhý) se nachází v rématu, označovat koreferenční vztah zdá jaksi divné. Avšak jaksi divné není žádný důvod ho neoznačovat a pokud se budeme při anotaci koreference opírat na jiné prostředky textové koheze, bude z toho kruh. Z toho plyne, že bychom měli označovat koreferenční vztah ve párech typu (41)-(42) za specifickou koreferenci. Srov.:

(41/In94204_107) Příjemně ji překvapilo , že se přihlásilo tolik dobrovolníků, kteří chtějí pomáhat druhým lidem.

(42/In94204_107) Nyní má linka třicet osm tzv. volontérů {coref_text, typ_SYN na „dobrovolník“}, kteří budou naslouchat volajícím.

??? K otázce výběru antecedentů u řetězců se specifickou referencí:

Situace, když entity se specifickou referencí se jednou vyskytují jako jednotlivé jednotky, jednou jako část celku a pak zase jednotlivě. V „parlamentním textu“ (In94210_95.t.gz) je to případ dvou paláců – Šternberský a Smiřických – na které se občas odkazuje zároveň a občas na jednotlivé. Srov.:

(21/In94210_95) Proto poslanci zhruba před rokem a půl rozhodli , že pro svou činnost znovu využijí blok renesančních paláců a domů mezi starou sněmovnou a Malostranským náměstím.... [...]

(23/In94210_95) Malostranské veřejnosti se však nápad poslanců příliš nelíbil : [...] paláce Šternberský a Smiřických (č . p . 6 a 7) i oba domy směrem do Tomášské ulice (č . p . 8 a 518) se staly sídly úřadů a normálními činžáky obývanými nájemníky .

(32/In94210_95) Paláce neznamenají přepych

(40/In94210_95) První patro Šternberského paláce skýtá ovšem přeci jen jednu výhodu : terasu vhodnou ke slunění (eventuálně k politickému řečnění) , protože je obrácená k jihu na Malostranské náměstí .

(42/In94210_95) Kromě pracoven bude v palácových patrech několik kuloárových chodeb... [...]

(56/In94210_95) Pražský magistrát pronajme soukromníkům obchody v podloubích obou šlechtických paláců , která směřují do Malostranského náměstí . [...]

(57/In94210_95) V nádvoří paláce Smiřických (možnosti - na 23jako ident. a na 56 jako SET. Tady je lepší na 56) by měla být zřízena dokonce kavárna pro veřejnost , kde by vzhledem k nižšímu podnikatelskému nájmu mohly být i nižší ceny než v okolí , tedy káva za [bůra]bůra místo za třicet .

V řetězci odkazů Šternberský palác - oba paláce - Šternberský palác, má poslední NP (Šternberský palác) odkazovat k NP(oba paláce) jako bridging, typ SET, nebo k předcházející NP(Šternberský palác) jako identická reference? Z hlediska textové koherence to může být různě v různých textech a kontextech. S docela velkou pravděpodobností je bridging-odkaz často logičtější. Odkazem spíše na poslední výskyt (i když to bude bridging) se zachová aspoň vzdáleně podoba řetězce. Ztrácí se však identická koreference mezi koreferenčními NP. Možná nejlépe dělat obě? Nebo podle smyslu...

Mám to v korpusu oanotováno dost chaoticky. Dělal jsem to spíš podle smyslu – zajímavé by bylo podívat se na vizualizaci a zjistit, jak se tam ta chaotičnost projevuje.

B.2.2.4. Koreference nereferenčních a generických NP (NR).

U nereferenčních NP nezaznamenáváme, je-li anaforická NP synonymum nebo hyperonymum antecedenta. Jinak by se to hodně zkomplikovalo. Opakující se v textu NP s nereferenčním statusem často může být synonymní nebo hyperonymní ve vztahu k předchozímu výskytu, takže pak by se nám ty charakteristiky množily. Například, vztah mezi mládež a děti v následujícím příkladě by měl dvě poznámky – NR a SYN:

(37/In94204_107) Na telefonní číslo 855 44 33 bude jistě volat mládež s různými problémy .

(38/In94204_107) Doufejme , že linka si časem vydobude mezi dětmi {coref_text, typ_NR na „mládež“} takovou autoritu , aby se na ni obracely i ty , které jsou skutečně ohrožovány .

Informace o synonymii nebo hyperonymii nereferenčních NP se mi nezdá ani natolik důležitá – stává se to docela často a na strukturu textu zdánlivě nebude mít vliv. U jmen se specifickou referencí očekávám pravděpodobně jiné chování synonymních a hyperonymních koreferujících členů – častější užívání ukazovacích zájmen, možné jiné pozice ve větě (vždy výrazně tématické) apod.

??? Hraniční případy u koreference NP se specifickou referencí a NP s textovou koreferencí typu NR.

V některých případech je těžko odlišit specifickou referenci od nereferenčních NP. Většinou možné jsou obě interpretace. Srov.:

(13/In9413_006) Po vojně začal v Masokombinátu v Ostravě - Martinově. [...]

(21/In9413_006) Ve státním podniku {coref_text, typ_NR nebo 0 na „masokombinát“} mne ubíjel stereotyp a nepružnost . [...]

(36/In9413_006) * A samotný start po odchodu z Martinova {coref_text, typ_SYN na „podnik“ nebo na „masokombinát“, pokud NP „podnik“ není označena jako koreferenční s „masokombinát“}? [...]
 (57/In9413_006) Klidně jsem mohl seskočit a dál dělat ve státním podniku {coref_text, typ_NR nebo 0 na „masokombinát“, „Martinov“ nebo na „podnik“}, nic by se nestalo .

Jmenné fráze v (21) a (57) referují stejně jako NP ve větách (13) a (36), ale mohou být pojaty i jako nereferenční, nebo spíše generické. Má cenu je odkazovat textovou koreferencí? Pokud ano, jako NR nebo jako specifickou koreferenci s poznámkou SYN nebo žádnou? Dále vidíme, že každá následující věta má o jeden směr šípky více. Variant je tady několik:

- pojmout NP „státní podnik“ v (21) jako nereferenční a neodkazovat ji na „masokombinát“. Pak s ní můžeme NR-vztahem propojit „státní podnik“ v (57), čímž vzniknou dva koreferenční páry (13) - (36) a (21) - (57)
- spojit mezi sebou všechny čtyři věty do jednoho koreferenčního řetězce se specifickou koreferencí, s poznámkou SYN.
- neoznačovat žádnou koreferenci u (21) a (57).

Ten problém zatím vyřešit neumím.

??? Další problematický bod v rozlišování specifické a nespecifické reference jsou **abstraktní jména**. Řešení je většinou dost arbitrární. Srov. např.

(32/In94204_107) Přiznal, z čeho má strach. [...]
 (35/In94204_107) Všechno nakonec dobře dopadlo, ale tohle dítě zbytečně prožilo půl roku strachu {coref_text, typ_0 na „strach“ nebo žádná šípka} a děsivých představ.

Zdá se moc divné propojovat tyto NP textovou koreferencí. Na druhou stranu je to pořád stejný strach. NEVÍM!!!! Je to celé problém a vůbec nevím, jak by se dal vyřešit... V daném příkladě má antecedent ještě navíc funktor CPHR, což také „ochuzuje“ jeho referenční status.

??? Vztah mezi **dějovými jmény** se zdá občas (netvrdím, že nemohou nikdy krásně koreferovat) ještě méně „koreferenční“. V případech (46)-(55) bych už koreferenci neoznačovala. Srov.:

(46/In94204_107) Linka 855 44 33 bude v provozu nepřetržitě 24 hodin. [...]
 (55/In94204_107) V těchto obdobích bude provoz {coref_text, typ_0 na „provoz“ nebo žádná šípka} na Lince bezpečí zněkolikanásoben , objasňuje ředitelka linky.

??? **Hraniční případy koreference s typem NR a něčím, co nemusí být jako koreference označeno**

Takových případů je velké množství. Je to ještě nereferenční koreference nebo už nic? Zatím to neumím elegantně vyřešit. V následujícím případě jsem to jako koreferenci neoznačila. Srov.:

(45/In94204_107) Kdy děti nejvíce volají [...]
 (48/In94204_107) Pražské děti budou mít hovor zdarma .

(49/In94204_107) Podle zkušeností ze zahraničí se dá předpokládat , že největší frekvence telefonátů nastane vždy mezi 16 . až 18 . hodinou .

Další příklady se stejným problémem:

(76/In94204_107) Chceme , aby ze sebe problémy dostaly ven .
(77/In94204_107) Jakmile se začnou svěřovat , už se s tím dá něco dělat .

(78/In94204_107) Jde o to , aby se z problémku nestal problém .

Neoznačovala bych (aspoň v poslední větě) ten vztah za koreferenční. Ale kde je ta hranice???

(13/In9413_006) Po vojně začal v Masokombinátu v Ostravě - Martinově .

(14/In9413_006) V dopravě , ale zajímal se o všechno z provozu . [...]

(24/In9413_006) Třeba při rozvozu jsem denně přenesl pěkných pár tun na zádech .

Textová koreference označená NR? Mohli bychom to k něčemu potřebovat? Těžko zachytitelné a nevypadá, že by mohlo mít pro strukturu textu nějaký důležitý význam. Ale zase - kde je ta hranice?

(22/In94210_95) Svědkem oněch časů zůstal mj . i pseudorenesanční spojovací můstek mezi sněmovnou a Šternberským palácem z roku 1910 .

(32/In94210_95) Paláce neznamenaají přepych

(36/In94210_95) Ač se to na první pohled nezdá , obývání klasických renesančních a barokních paláců {??? s velikými , řetězovitě propojenými místnostmi není žádné terno .

Tady máme NP s generickou referencí (klasických renesančních a barokních paláců), která se už předtím v textu několikrát vyskytla, ale žádný odkaz, prostě generická NP. V některých jiných případech jsem však označovala vztahy mezi generickými NP v textu jako koreferenční (značka NR). Kde je hranice?!

Šlo by tu nabídnout, že pokud „antecedent“ má výrazně specifickou referenci, koreferenci neoznačovat. Pokud je to vztah mezi dvěma NR, a v textu se na to navazuje, tak označovat s poznámkou „NR“. Ale chtělo by se to ještě ověřit. Zrovna v tomto případě odkaz by byl příp. na (32), kde NP paláce ani nemusí mít specifickou koreferenci, i když odkazuje na předchozí věty, kde jsou konkrétní (*Šternberský a Smiřický*) paláce.

(4/In94210_95) Když se koncem osmdesátých let rekonstruovala historická budova českého sněmu na Malé Straně v Praze , nikdo z architektů asi nepředpokládal , že už za pět let bude svým uživatelům malá . [...]

(64/In94210_95) Když architekti zvažovali optimální propojení staré budovy sněmovny s novými domy , vsadili na tunel pod Thunovskou uličkou .

Nejraději bych to nijak neoznačovala, ale musím na to mít kritéria

Sporné je označování textové koreference u nereferenčních NP, opakujících v textu na velkou vzdálenost (dejme tomu přes 5 vět), v případě, když nejsou žádné známky koherence. Srov.:

(40/ In94204_107) Když si dítě bude přát , aby se o jeho problému nikdo z rodiny nebo školy nedozvěděl , musíme to respektovat , vysvětluje Jana Drtilová . [...]

(62/ In94204_107) Linka by neměla rodinu {??? coref_text, typ_NR na „rodina“} nahrazovat , ale doplňovat .

V podobných případech nabízím označovat koreferenci se známkou NP jenom na krátkou vzdálenost (poslední výskyt nesmí být dříve, než před 5 větami a také v případech, že na koreferenci poukazují jiné kohezní prostředky – ukazovací zájmeno, jasná kontextová zapojenost apod.) Takové řešení však není moc elegantní a je možná zavádějící.

Další vztah je na dost velkou textovou vzdálenost, avšak tady už možná něco koherentního je:

(49/In94210_95) Poslanci budou muset odpočívat jinde , protože suterény jsou příliš hluboko a napojení na výše položenou kanalizaci pomocí čerpadel by provoz budov neúměrně prodražilo .

(66/In94210_95) Existující kanalizační sítě by totiž podzemní chodbu vtlačily tak hluboko do země , že by jí jistě nikdo nepoužíval .

C. Bridging vztahy

Bridging vztahy označujeme u NP, které nejsou koreferenční, ale mezi nimi existuje určitý sémantický vztah. Určitý neznamená jakýkoliv – označujeme jenom několik málo, a tento výběr je založen na literatuře z oblasti computer linguistics. (pokud bude třeba, teoretický výklad pak napíšu podrobněji)

Pozor! Při anotaci bridging vztahů musíme dbát na to, že elementy, mezi kterými postulujeme vztah, mohou být lexikálně stejně vyjádřeny, ale přitom nebýt koreferenční. A naopak, povrchová textová realizace párů může navádět na vztah typu bridging, zatímco ve skutečnosti jde o identickou (textovou) koreferenci. Srov např.:

(38/ In94204_107) Doufejme , že linka si časem vydobyde mezi děťmi takovou autoritu , aby se na ni obracely i ty , které jsou skutečně ohrožovány .

(40/ In94204_107) Když si dítě {bridging, typ_SET na „dítě“} bude přát , aby se o jeho problému nikdo z rodiny nebo školy nedozvěděl , musíme to respektovat , vysvětluje Jana Drtilová .

(100/In9413_006) Určitou svou představu si chci ověřit a potvrdit při cestě po USA , kterou jsem obdržel za vítězství y soutěži podnikatelů {má specifickou referenci}.

(101/In9413_006) K různým soutěžím {bridging, typ=contrast, na „soutěž“} mám výhrady .

(34/ In94204_107) Péťa skončil u Jany Drtilové.

(35/ In94204_107) Všechno nakonec dobře dopadlo , ale tohle dítě {coref_text, typ_ER na „Péťa“} zbytečně prožilo půl roku strachu a děsivých představ .

Pozor! Pomocí bridging-vztahů propojujeme (pokud možno) jenom lexikálně vyjádřené autosémantické uzly (žádné „který“ a rekonstruovaný „PersPron“). V dnešních anotovaných datech jsem to však nedodržovala pravidelně – jsou to chyby, které opravím. Nebo případy typu (24/ In9413_006), kde to jinak zatím nedokážu vyřešit.

V rámci anotace bridging-anafory vyčleňujeme čtyři základní vztahy a jednu sběrnou pro to, co zatím nemůžeme nikam zařadit. Jsou to:

- | | |
|-----------------|--|
| PART | – vztah část-celek (oba směry), viz C.1. |
| SET | – vztah množina-podmnožina (element množiny) (oba směry), viz C.2. |
| FUNCT | – vztah individual – function, viz C.3. |
| CONTRAST | – vztah sémantického protikladu, viz C.4. |
| REST | – blíže neupřesněná kategorie: je vyznačený „bridging“ vztah, ale není specifikovaný jeho druh, viz C.5. |

C.1. bridging-vztah část-celek (PART)

Vztah části a celku je asi základní bridging vztah, který se nejčastěji vyskytuje. Vzorové příklady: *pokoj-strop*, *ruka-prst* apod. Podle posledních anotovaných dat, vypadá, že ten vztah může být oboustranný, tj. vztah *celek-část* je stejně přínosný a důležitý pro koherenci jako *část-celek*. Srov. např. oba směry vztahu, reprezentované v následující větě:

(42/In94210_95) Kromě pracoven {bridging, typ=PART, na „patro“} bude v palácových patrech několik kuloárových chodeb {bridging, typ=PART, na „patro“}, zasedacích místností {bridging, typ=PART, na „patro“} a přijímacích salonků {bridging, typ=PART, na „patro“} s barokními stolky {bridging, typ=PART, na „salonek“} a křesly {bridging, typ=PART, na „salonek“}.

!!! V případě, když máme vztah část-celek v jedné větě, přičemž celek je ZA částí, a přitom se ještě někde v předchozím, ne bezprostředně blízkém kontextu vyskytuje další možný celek, spíše tu část odkážeme na celek, který je vepředu. Srov.:

(45/In94210_95) Žádné honosné taneční sály nebo restaurace v novém komplexu nebudou . [...]
(49/In94210_95) Poslanci budou muset odpočívat jinde , protože suterény (bridging, typ=PART na „budova“) jsou příliš hluboko a napojení na výše položenou kanalizaci pomocí čerpadel by provoz budov neúměrně prodražilo.

??? U příkladů typu (42) je logické položit otázku, zda li je smysluplné rozlišovat tolik sémantických vztahů uvnitř jedné věty, kde už samotná syntaktická struktura částečně tyto vztahy naznačuje (předložka „s“ a funktoři ACMP u NP stolek a křeslo). V dané konkrétní větě z hlediska

textové koherence to není nezbytně nutné. Na druhé straně u jiné podobné věty to může být docela důležité. Řešení na to zatím nemám.

Sporný příklad 1 (odkaz k něvyjádřenému členu):

(24/In9413_006) Třeba při rozvozu #PersPron jsem denně přenesl pěkných pár tun na zádech {bridging, typ=PART, na „#PersPron“}.

Koreferující člen odkazuje k nevyjádřenému zájmenu – inkonsistence, o které jsem už zmínila výše, že občas odkazování u bridging jsem nedělala pravidelně (viz ...). Ale protože je to přímá řeč a explicitně vyjádřená první osoba se ani jednou nevyskytuje, tak ani nevím, jak bych to jinak řešila.

Sporný příklad 2 (velká vzdálenost mezi členy (28 vět) a velice sporná sémantická souvislost – může být jenom náhoda):

(5/In9413_006) Tak hodnotí Petr Chodura , podnikatel z Ostravy , první momenty po oznámení , že se stal Vynikajícím podnikatelem roku 1993 .

(33/In9413_006) Nyní mám firmu , která vyrábí více jak dvě tuny masných specialit denně , mám pět obchodů na severu Moravy {bridging, typ=PART, na „Ostrava“} .

NP na severu Moravy ve větě (33) je navíc v rématu, takže nemusí v každém případě odkazovat k předchozímu kontextu. Avšak pokud tuto návaznost přece jenom postulujeme, taková informace nemusí být úplně sémanticky prázdná. (Třeba už ze začátku textu víme, že tento podnikatel je z Ostravy a teď ji rozšíříme na celou severní Moravu – ale zní to všechno až moc vágně a trochu zoufale).

Další příklady, normální:

(44/In9413_006) Dělal jsem bez přestávky celé týdny , často v nocí {bridging, typ=PART, na „týden“} .

(9/In94210_95) [...] budova ČNR praskala ve švech , ovšem do roku 1992 to zajímalo jen málokoho. [...]

(15/In94210_95) Pokud tedy zrovna nesedí na svém minikřesle v jednací síni {bridging, typ=PART, na „budova“}, jsou poslanci nuceni pobývat buď ve svých klubech {bridging, typ=PART, na „budova“}, nebo postávat či posedávat po chodbách {bridging, typ=PART, na „budova“} .

C.2. bridging-vztah množina-podmnožina/element množiny (SET)

Vztah mezi množinou a podmnožinou nebo elementem této množiny. Vzorové příklady: *Mušketýři – Athos - Porthos - Aramis; nápoje – pivo – limonáda – minerálka – cola; motýli – červení – bílí; semináře – první seminář – poslední seminář*

Stejně jako u vztahu část-celek vztah SET může být oboustranný, tj. že vztah *element-množina* není chudší ani méně reprezentativní, než *množina-element*. Srov. např.:

(14/In94210_95) Na rozdíl od dobře vybaveného FS dnes nikdo z téměř dvou stovek poslanců kromě předsedy {bridging, typ=SET, na „poslanec“} a místopředsedů {bridging, typ=SET, na „poslanec“} sněmovny a šéfů {bridging, typ=SET, na „poslanec“} jejích výborů nemá svou kancelář , pracovní stůl , židli a telefon .

(15/In94210_95) Pokud tedy zrovna nesedí na svém minikřesle v jednací síni , jsou poslanci nuceni pobývat buď ve svých klubech, nebo postávat či posedávat po chodbách .

(16/In94210_95) Nelze se pak ani divit , že část zákonodárců {bridging, typ=SET, na „poslanec“} zvolí příjemnější variantu a odchází úřadovat do suterénní restaurace zvané dolní sněmovna .

Další příklady:

(33/In9413_006) Nyní mám firmu , která vyrábí více jak dvě tuny masných specialit denně, mám pět obchodů na severu Moravy .

(62/In9413_006) Firma produkuje na padesát sortimentních druhů párků {bridging, typ=SET, na „specialita“}, klobásek {bridging, typ=SET, na „specialita“}, salámů {bridging, typ=SET, na „specialita“}, vyjma trvanlivých.

(17/ In94204_107) Na toto telefonní číslo však mohou samozřejmě zavolat všichni kluci a děvčata , kteří se ocitnou ve svízelné situaci .

(18/ In94204_107) Ptali jsme se několika {bridging, typ=SET, na „(kluci a (děvčata))“}, jestli by takového kamaráda po telefonu považovali za dobrou věc.

Sporný příklad 1 (vztah SET uvnitř jedné věty).

V případě identické textové koreference jsme se dohodli na tom, že nebudeme označovat koreferenční vztah mezi subjektem a predikátovou částí výpovědi. Je správné zachovat toto řešení v případě bridging anafory, kde vztah mezi subjektivou a predikátovou částí výpovědi už není takový triviální? Na existenci a typ vztahu sice může částečně poukazovat syntaktická a lexikální struktura, nedává to však vyčerpávající informaci. Srov.:

(43/In94204_107) Jsou mezi nimi například studenti vysokých škol {bridging, typ=SET, na „#PersPron“}, herečka {bridging, typ=SET, na „#PersPron“}, kunsthistorik {bridging, typ=SET, na „#PersPron“}, učitelka {bridging, typ=SET, na „#PersPron“}, psycholožka {bridging, typ=SET, na „#PersPron“}.

Zatím jsem to udělala jako SET, ale nejsem si tím úplně jistá.

Sporný příklad 2 (otázka nutnosti a hloubky interpretace).

Je v následujícím příkladu vztah SET?

(44/ In94204_107) [volontéři] Absolvovali školení v první pomoci pro člověka v nouzi . [...]
 (56/ In94204_107) Když dítě zavolá , dostane buď radu hned , nebo si s ním volontér domluví další hovor .

Sporný příklad 3 (označování některých bridging vztahů u nereferenčních NP)

Srov. příklad:

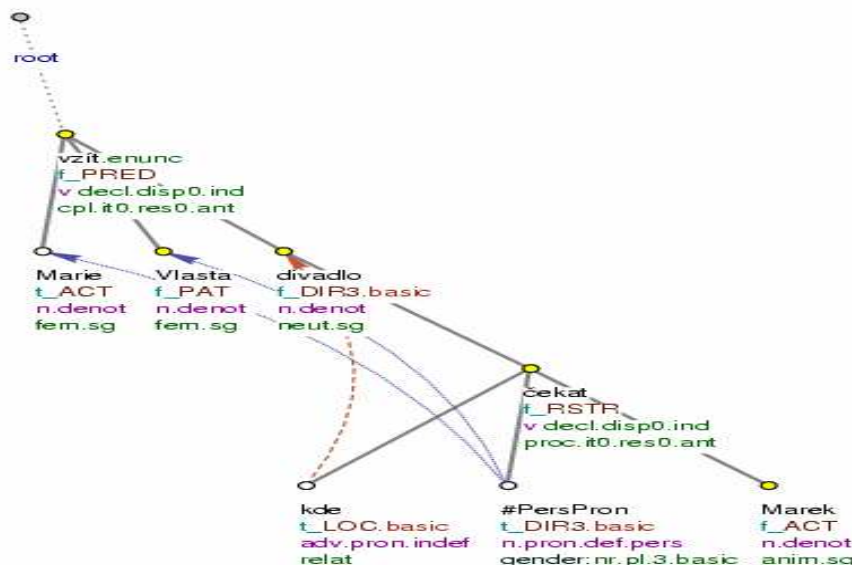
(51/In94210_95) Ani pro parlamentní knihovnu nezbude na Malé Straně místo .

(53/In94210_95) Umístit knížky na speciálně upravených půdách , jako je tomu v Národní knihovně v Klementinu , prý z bezpečnostních důvodů nelze .

Vztah mezi knihovna a knížky je sice SET, ale k čemu na kdy může být dobré? Označovat nebo neoznačovat? A takových příkladů je velké množství. Ani to není nabídka, ale hodně abstraktní úvaha – že by mohly NR spíše zanedbávat, zatímco KR označovat. Ale nevím....

Odkazování k více uzlům

!!! Inkonsistence vzniká při srovnání původní textové koreference s bridging anaforou typu SET a PART v případech, kdy vyjádřeny zájmenem koreferující uzel odkazuje k dvěma uzlům v předcházejícím kontextu. V anotaci PDT 2.0. se dohodlo na prozatímním technickém řešení – k oběma uzlům je nutno odkázat jednotlivě.



Srov.:

Marie vzala Vlastu do divadla, kde na ně {coref_text, na „Marie“ a „Vlasta“} čekal Marek.

V rozšířené anotaci se však setkáváme s větami, kde to nebylo v předchozí anotaci označeno. Logické by bylo v těchto zatím neoanotovaných případech udělat bridging s typem SET (možná občas PART). Ale copak dělat s těmi, co jsou už udělány? Srov. nové věty:

(23/In94210_95) [...] paláce Šternberský a Smiřických i oba domy směrem do Tomášské ulice se staly sídly úřadů a normálními činžáky obývanými nájemníky .

(24/In94210_95) Ti {bridging, typ=SET, odkazy na „úřad“ a „nájemník“} se nyní měli vystěhovat a někteří občané to pochopili jako znárodňovací akt.

V anotaci jsem podobné trojice propojovala jako bridging, typ_SET.

??? Hraniční případy mezi bridging typy SET a PART.

Když jsem za sebou kontrolovala anotaci třech zkušebních souborů, zjistila jsem, že se mi až nepřiměřeně často pletou typy SET a PART. Na některých místech je to oprávněno více, na některých méně, docela málo se najde čistých příkladů na PART, a ještě aby ten vztah plnil nějakou koherenční funkci. Zatím nechci tím nic tvrdit, ani likvidovat vztah PART, musíme se podívat na větší počet textů a s nimi související statistiky. Avšak jak se zdá podle prací spřátelených kolegů, jenom málo pracovišť tyto vztahy mezi sebou rozlišuje.

Srov. např.:

(47/In94210_95) Jeho hlavní výhodou by mělo být lepší napojení na televizní přenosovou techniku : zatímco dnes přenosové vozy {bridging, typ=SET nebo PART, na „technika“} blokuji parkovací prostor před starou sněmovnou , v budoucnu zajedou do Thunovské a kabely {bridging, typ=SET nebo PART, na „technika“} se snadno spojí s tiskovým centrem .

V daném případě v podstatě řeším jenom, jestli je jméno počitatelné nebo není. Pokud ano, pak SET, pokud ne, tak PART. Nemělo by tak být, kritéria by měla být založena na něčem jiném.

C.3. bridging-vztah sémantického protikladu (CONTRAST)

Vztah sémantického protikladu také ve velké míře přispívá ke koherenci textu, proto jsme se rozhodli ho označovat v anotaci.

Tato skupina se terminologicky prolíná s kontrastem v AČ. Částečně i významově. Jak ukazují příklady, některé bridging-šípky s poznámkou CONTRAST mají c-čko i v anotaci AČ. Není to však pravidlem. Srov. příklady:

V anotaci AČV oba členy vztahu mají f(okus):

(53/In9413_006) A přesvědčen jsem ještě o jednom - je třeba mít vysoké cíle a s malými [cíly] {bridging, typ=CONTRAST, na „cíl“} se nespokojit .

V anotaci AČV oba členy vztahu mají c(ontrast):

(21/In94204_107) Co se může dospělému zdát zanedbatelnou záležitostí, naroste v dětské {bridging, typ=CONTRAST, na „dospělý“}mysli třeba i do tragických rozměrů.

V anotaci AČV oba členy vztahu mají t(opic):

(6/In94210_95) Její poslanci se před rokem 1989 scházeli čtyřikrát do roka , odhlasovali vše , co se jim řeklo , a pak se rychle vrátili do svých domovů , kde některým běžel plat koncernových ředitelů a jiným dojiček krav .

(8/In94210_95) Po listopadu 89 {bridging, typ=CONTRAST, na „rok“} se poslanečství stalo placenou činností a nároky na jeho vykonávání přiměřeně tomu vzrostly.

C.4. bridging-vztah zatím neterminovaný (REST)

blíže neupřesněná kategorie: je vyznačený „bridging“ vztah, ale není specifikovaný jeho druh.

Varianty: *ctitel – obdiv, trůn – král, duševní stav – profesní psychiatr – diagnóza, kalhoty – kůže.*

Srov. např.:

(10/In9413_006) Úzce navazuje na tradici podnikání svého rodu, především dědy.

(12/In9413_006) Od něj {coref_text, typ_0 na „děda“} získal vnuč {bridging, typ_REST na „#PersPron“} výtečné základy , ač sám vystudoval školu zaměřenou na dopravu.

Pozor! Jak uvedeno výše, bridging šípka spojuje lexikálně (většinou podstatným jménem) vyjádřené uzly.

(26/In9413_006) Jak složitý byl přechod na podnikatelskou dráhu?

(28/In9413_006) Hledal jsem ten nejsprávnější směr {bridging, typ_REST na „dráha“}.

Mezi uvedenými dvěma příklady je rozdíl v kontextovém zapojení druhého členu páru. Zatímco v (12) NP vnuč je jednoznačně KZ, přičemž tato KZ je dána pozicí ve větě ale také i sémantickým vztahem děda-vnuč, ve větě (28) to není úplně jednoznačné. Otázka, jestli pro pár dráha-směr v tomto příkladě je smysluplné určovat bridging vztah, zůstává v podstatě otevřenou. Osobně bych se přiklápěla spíše k označování a v anotaci jsem to udělala, nevymyslela jsem však na to zatím žádná dobrá kritéria.

Další příklady:

(55/In9413_006) Začal jsem, řekněme, jako provazochodec .

(56/In9413_006) Lidé chodili po zemi , já nějakých dvacet centimetrů nad ní . (57/In9413_006) Klidně jsem mohl seskočit a dál dělat ve státním podniku , nic by se nestalo .

(58/In9413_006) Ale začal jsem lano {bridging, typ_REST na „provazochodec“} zvedat a seskočit už nebylo možné.

(58/ In94204_107) Pokud to bude potřeba a dítě k tomu dá souhlas , pozve je {coref_text, typ_0 na „dítě“} v doprovodu rodiče {bridging, typ_REST na „#PersPron“} nebo jiného dospělého do dětského krizového centra, jež tvoří zázemí linky.

Následující příklad je problematičtější – jestli se to má označovat ještě jako bridging, nebo už nemá, jak to neděláme u hypertématických vztahů (viz tady dále):

(40/ In94204_107) Když si dítě bude přát, aby se o jeho {coref_text, typ_0 na „dítě“} problému nikdo z rodiny {bridging, typ_REST na „dítě“} nebo školy nedozvěděl, musíme to respektovat, vysvětluje Jana Drtilová.

Ad hypertématické propojení textu.

Vztah C.4. má blízko k další kategorii vztahů, která však už přesahuje pojem bridging-anafory, tj. ke spojování prvků do společného sémantického pole a určování hypertématických řetězců textů. Vyhledávání hypertématických linků dává možnost uvidět základní linii (linie) článku. (Např. v cmpr9410_028.t. budou dvě – umění a peníze.) To může určitě k něčemu technickému přispět - srov. např. bohaté výzkumu googlu a jiných na to téma – dnes moc populární, třeba i k vyhledávání podle tématu (v nelingvistickém smyslu slova „téma“). Tady je příklad jednoho hypertématického řetězce z prvních 60 vět souboru cmpr9410_028.t.gz:

peníze – investování – cena – koupit – prodělat – prodat – pořídit – stát – zhodnocení – poptávka – trh – nabídka – kupce – rozpočet – hodnota – nejdražší – nejlevnější – 9 tisíc – zaplatíte – kupující – ceny stouply – trojnásobek – vyplatí se – prodej atd.

Na jedné stráně by bylo škoda ty výrazy vůbec nespojit. Identickou koreferencí je spojit všechny nemůžeme, protože referují vždy k jiným objektům. Zůstává bridging, ale těch málo typů, které jsou pro bridging uvedeny v literatuře, rozhodně nestačí, něco budeme muset vynechat, na něco nevymyslíme vztah apod. Hypertématický vztah by je tedy bez problému spojil. Na druhé stráně ruční provedení takové práce je velice časově náročné. Do takového řetězce mohou patřit všechny autosémantické lexikální prvky, bez zřetele k jejich slovnímu druhu a referenční platnosti. Občas, a to dost často, v jedné větě se najde několik uzlů daného sémantického pole, které budou mezi sebou propojeny. Tyto vztahy se pak zase složitým způsobem proplétají do existujících vztahů textové koreference a bridging. Kromě toho, v počítačové lingvistice existují i jiné než ruční způsoby vyhledávání hypertémat článků. Z těchto důvodů hypertématické vztahy v stávající anotaci neoznačujeme.

D. Speciální typy reference (coref_special)

K existujícím odkazům jsem dodala případy exophorického odkazu v případě když odkazovací výraz se neskládá jenom ze zájmena, ale obsahuje normální podstatné jméno. K takovým případům patří především časová (v *tomto roce*) ale i prostorová deixe. Srov. příklady:

Časová deixe se v anotovaných datech zatím neobjevila. Uvedu však příklad z SYN2005:

Po chvílce mlčení se Billy otázal: "Je tady už ten druhý?"
"Zatím ne." Ale právě v tom okamžiku se objevil muž, jehož tvář viděl Broderick v dokumentaci Teda Sanderse vedle fotografie Hubbardovy. (Polák, J., Závody)

Byl krásný den , první skutečně jarní den toho roku. (Polák, J., Závody)

Další příklady s mimojazykovými odkazy:

(28/In94210_95) Dokončeny by měly být do 31 . prosince 1995 ,
a to i přes jisté zdržení způsobené opožděným stěhováním
nájemníků z domů čp. 8 a 518 do náhradních bytů na sídlišti
Barrandov v těchto dnech {coref_special, typ exoph}.

Odkazů na segmenty textu jsem dodala jenom málo – většinou tam už byly, a nezaměřila jsem se na to.

Hraniční případy mezi typem coref_special, typ segm a bridging anaforou, typ SET.

Srov.:

(77/In9413_006) Spolupráce by měla dostat patřičný rytmus ,
režim .
(78/In9413_006) Vysoké využití podhorských pastvin ,
nejkvalitnější stáda .
(79/In9413_006) To jsou předpoklady pro výrobu kvalitních
potravin .

??? V dané větě UZ „to“ odkazuje v původní anotaci na širší předchozí kontext jako segm.
Pravděpodobně to tedy referuje k celé predikaci:

to = spolupráce má dostat patřičný rytmus, režim,

Druhá možná interpretace, je:

předpoklady = { patřičný rytmus, režim, vysoké využití
podhorských pastvin, nejkvalitnější stáda, ...}

Výběr je podle na anotátoru a smyslu věty. Sémantická ambiguita.

* * *

Ad named-entities.

Zvláštní pozornost má být věnována anotaci vztahů, kde aspoň jeden z páru je pojmenovaná entita. Anotace koreference pojmenovaných entit je věnována v současné počítačové lingvistice, existují fungující programy k její automatickému rozpoznávání. U nás teď pracuje skupinka na automatickém rozpoznávání a klasifikaci pojmenovaných entit. Tuto informaci bude třeba použít pro naši koreferenční analýzu. Je třeba však i v tomto případě rozlišovat mezi textovou koreferencí a vztahem typu bridging, kdy členy páru nejsou mezi sebou vzájemně koreferenční. Vztahy *Bělorusko – president* a *Lukašenko – president* nejsou stejné. V prvním případě to bude bridging-vztah s typem FUNCT, v druhém – identická textová nominální koreference.

Pojmenované entity se budou častěji opakovat bez identifikátoru. Srov. např.

(46/94210_95.t) Pouze z bývalé Šternberské konírny v přízemí
křídla přiléhajícího k Thunovské uličce se stane konferenční
(tiskový) sál. [...]

(64/94210_95.t) Když architekti zvažovali optimální propojení
staré budovy sněmovny s novými domy , vsadili na tunel pod
Thunovskou uličkou .

A bývají často nahrazovány jménem funkce (typ *Lukašenko – president*), ale ponechává přitom identická koreference.

Problémy a hodnocení předpokládané kvality budoucí anotace

Při příp. anotování rozšířené koreference „ve velkém“ vyvolá podle očekávání následující větší problémy:

- Prostředky textové koreference jsou svou povahou dost vágní, a pokud se neomezujeme jenom na zájmena, jak to bylo u základní anotace, uvidět všechny páry a správně je označit není úplně jednoduchý úkol.
- Velký problém vyvolá identická koreference nereferenčních NP, především samotný výběr těchto NP – co označovat a co neoznačovat a proč. Pokud v textu nejsou žádné další koherenční prostředky, výběr záleží v podstatě jenom na intuici, a to není dobře.
- u bridging-anafor, kromě vyhledávání správných párů, je problematické (ne)zařazení do typu REST – existence nejasného sémantického vztahu. Existuje dost pohraničních příkladů mezi SET a PART v případech nepočitatelných a abstraktních referentů.

Uvedené faktory pravděpodobně způsobí (aspoň na začátku) velkou anotátorskou neshodu, takže s tím budeme muset něco udělat.

Rovina analýzy diskurzu

Tektogramatická rovina je dost úplná a rozšířená anotace textové koreference a bridging vztahů je v podstatě už informace, kterou tam přinášíme navíc. Tato informace již přesahuje sémantiku věty a vztahuje k její pragmatice, referenci celého textu a jeho komponentů ke skutečnosti apod. V současné době se uvažuje o zavedení hlubší roviny, tj. roviny diskurzivní, ale její podoba ještě není dostatečně definována. Pokud však taková rovina bude zavedena, bude pravděpodobně logické přesunout tam i naši rozšířenou anotaci koreference.

Statistiky:

Anotováno dat:

	vět	uzlů
In9413_006 (Podnikatel roku)	128	1235
In94204_107 (Kamarád po telefonu)	103	1258
In94210_95.t.gz (Parlament rozšiřuje své sídlo)	105	2269
<i>Celkem</i>	336	4762

Statistika počtu vztahů:

		In94210_95.t.gz	In9413_006	In94204_107	Celkem
A. gramatická koreference	<i>Celkem gramatické</i>	52	36	41	129
B. textová koreference identická	NR	49	28	71	148
	SYN	17	3	5	25
	ER	0	1	5	6
	0 = zbytek	126	58	74	258
	<i>Celkem identické</i>	192	90	155	437
C. bridging anaforické (kataforické) vztahy	PART	26	28	1	55
	SET	29	3	22	54
	FUNCT	2	3	2	7
	CONTRAST	13	5	11	29
	REST	2	5	6	13
	<i>Celkem bridging</i>	72	44	42	158
D. coref_special	exoph	1	0	1	2
	segm	1	5	2	8
	<i>Celkem coref_special</i>	2	5	3	10
<i>Celkem</i>		318	170	238	734

Srovnání s předchozí verzí od Lucky:

A. gramatická koreference	<i>Celkem gramatické</i>	129
B. textová koreference identická	<i>Celkem identické</i>	275
C. bridging anaforické (kataforické) vztahy	<i>Celkem bridging</i>	9
<i>Celkem</i>		413

Seznamy typů s příklady:

bude doplněno