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 brigding vztahů v Pražském závislostním korpusu.

Anja Nědolužko

1. září 2008

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

ÚVOD....6

1. OBECNÉ POZNÁMKY....6

X.I. Cíle anotace....6

X. PŘEHLED LITERATURY K TÉMATU....7

X.1. Literatura k teorii reference....7
X.X.X.X. E.V. Padučeva....7
X.X.X.X. Šmelev1996....9
X.X.X.X. B. Palek1988....10
X.X.X.X. Yokoyma....12
X.X.X.X. Mendoza2004....12
Řešení....12
X.X.X. Predikace vs. Identifikace....14

X.2. Anotace – literatura z oblasti zpracování koreference v počitačové lingvistice....18
Chiarcos, Krasavina 2005....18
Vieira – Teufel 1997....18
MATE....19
MATE Dialogue Annotation Guidelines,....19
Massimo Poesio, 2004. "The MATE/GNOME Scheme for Anaphoric Annotation, Revisited", Proc. of SIGDIAL, Boston, April....20
MUC....21

A. GRAMATICKÁ KOREFERENCE....27

A.1. Dodržování koreferenčního řetězce....27

B. TEXTOVÁ KOREFERENCE....28
Textová koreference na velkou textovou vzdálenost....28

B.1. Původní zájmenná koreference – některé případy pronominalizace a elipsa....30

B.2. Rozšířená anotace textové koreference....30
B.2.1. Slovnědruhová charakteristika koreferovaných párů....30
B.2.2. Typologie textově koreferenčních vztahů....32
B.2.2.1. Koreferenční vztah mezi NP se specifickou referencí, kde koreferující člen není synonymum ani hyperonymum antecedenta....33
Kataforický odkaz dopředu....34
Koreference osobních zájmen v dialogických textech....34
Koreference otázkového slova a odpovědi v dialogických textech....35
B.2.2.2. Koreference synonymních NP (SYN)....35
B.2.2.3. Koreference hyponymu a hyperonymu (ER)....37
B.2.2.4. Koreference nereferenčních a generických NP (NR)....38
??? Hraniční případy u koreference NP se specifickou referencí a NP s textovou koreferencí typu NR....39
B.2.3. Problématické případy označování textové koreference....40

B.2.3.1. Hraniční případy koreference s typem NR a něčím, co nemusí být jako koreference označeno....40
B.2.3.2. Koreference abstraktních jmen....42
Přehled literatury k tématu....42
Literatura k řešení:....43
B.2.3.3. Koreference dějových jmen....44
B.2.3.4. Problematické páry NP se specifickou referencí....44
Sporný příklad (koreference specifické a autonymní NP)....45
B.2.3.5. Dvě místní určení vedle sebe (tady v Praze, u nas doma apod.)....45
Lingvistická zajímavost – anafora na neobligatorní a nevyjádřené místní určení:....45
B.2.4. Nejednoznačný výběr antecedentu....45
B.2.4.1. K otázce výběru antecedenta v případe apoziční skupiny:....45
B.2.4.1. K otázce výběru antecedenta v případe koordinační skupiny:....46
B.2.4.3. K otázce dvojího odkazování (identická koreference) - ???:....46
B.2.4.4. Spojení se slovy s funkcí „kontejneru“....47

C. BRIDGING VZTAHY....47
Přehled literatury k tématu bridging....47

C.1. Typologie bridging vztahů....48
C.1.1. bridging-vztah individual – function (FUNCT)....49
Bridging vztah typu FUNCT v páru ministr Karel Dyba - ministerstvo....49
Kontextově podmíněný vztah FUNCT....49
C.1.2. bridging-vztah část-celek (PART)....50
C.1.3. bridging-vztah množina-podmnožina/element množiny (SET)....52
Sporný příklad 1 (vztah SET uvnitř jedné věty).....52
Sporný příklad 2 (otázka nutnosti a hloubky interpretace).....53
Sporný příklad 3 (označovaní některých bridging vztahů u nereferenčních NP)....53
??? Hraniční případy mezi bridging typy SET a PART....53
C.1.4. bridging-vztah sémantického protikladu (CONTRAST)....55
C.1. 5. bridging-vztah zatím neterminovaný (REST)....56

C.2. Skupiny bridging vztahů....57
C.2.1. Vztah „místo – obyvatel“.....57
C.2.2. Vztah typu „autor – kniha“....57
C.2.3. Vztah „věc – majitel“....57
C.2.4. Vztah mezi stejně vyjádřenými nebo synonymními nekoreferenčními NP....58
C.2.5. Vztah událost – argument....59
C.2.6. Anafora bez koreference....59

C.3. Nejednoznačný výběr antecedentů....59
C.3.1. Spojení se slovy s funkcí „kontejneru“....59
C.3.2. K otázce výběru antecedentu v případe apoziční skupiny:....60
C.3.2. K otázce výběru antecedentu v případe koordinační skupiny:....60

D. Bridging nebo textová identická – výběr a preference....60
D.1. K otázce dvojího odkazování (textový a bridging vztahy) a preference:....61
D.2.Odkazování k více uzlům....61
D.3. K otázce výběru antecedentů u několika sémantický spojených řetězců se specifickou referencí (textová + bridging):....63
D.3.1. Dlouhé mezi sebou propojené řetězce....63

D.3.2. „faktory – jeden z faktorů“ 64
Případ „zaměstnanci – každý ze zaměstnanců“ 65
Připustitelná nepřesnost v určování bridging vztahů 65
Ad hypertématické propojení textu. 66

D. SPECIÁLNÍ TYPY REFERENCE (COREF_SPECIAL) 66

Exoforické odkazy 66

Odkazy na segmenty textu 67
Hraniční případy mezi typem coref_special, typ segm a bridging anaforou, typ SET. 68
Ad named-entities. 69
Koreference u spojení obecného jména a pojmenované entity 69
Anotace částí pojmenovaných entit 70

MOŽNOSTI AUTOMATIZACE ROZŠÍŘENÉ ANOTACE KOREFERENCE A BRIDGING VZTAHŮ 71

Co má být doděláno automaticky: 71

PROBLÉMY A HODNOCENÍ PŘEDPOKLÁDANÉ KVALITY BUDOUCÍ ANOTACE 71

ROVINA ANALÝZY DISKURZU 72

TECHNICKÉ PROBLÉMY KOMBINACE ROZŠÍŘENÉ A PŮVODNÍ ANOTACE KOREFERENCE 72

PŘÍLOHA X: SROVNÁNÍ KONCEPCE PDT A JINÝCH PŘÍSTUPŮ 77

Chiarcos, Krasavina 2005 77

REFERENCES: 78

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. Anotace koreference je představena ve „velkém manuálu“ (Mikulová a kol. 2005). Podrobný popis anotačního schématu,a to jak po stránce lingvistické, tak po stránce technické, je obsažen v (Kučová a kol. 2003)

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

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.

1. Obecné poznámky

Momentálně existující anotace substantivní koreference a asociační anafory – buď velice obecná kritéria s lepší mezianotátorskou shodou (Hirschman 1998) nebo naopak příliš specifická kriteria, která jsou špatně automaticky zpracovatelná.
V anotaci PDT jsme se pokusili o kompromis – na jedné straně základní typy, které se dají zpracovat přesnými metody, na druhé straně však dost podrobná sémantická klasifikace.

Anotaci koreference provádíme na syntakticky oanotovaných tektogramatických stromech PDT. Výhodou toho je, že některé informace o koreferenci jsou již zahrnuty do struktury stromu. Tak např. nemusíme spojovat uzly, které jsou mezi sebou ve vztahu apozice nebo predikace. Podrobněji _VIZ_

X.I. Cíle anotace

Následující text popisuje rozšířenou anotaci koreference na velkém korpusu textů. to může sloužit k řešení následujících úkolů:

- Pro teoretický výzkum
    - vlastnosti koreferenčních a anaforických výrazů
    - použití NP v anaforické pozici
    - saliences apod.
- Pro praktické aplikace:
    - automatická generace referenčních výrazů
    - anaphora resolution
    - evaluace
    - automatické porozumění textu
    - machine learning apod.

2. Přehled literatury k tématu

X.1. Literatura k teorii reference

X.X.X. E.V. Padučeva

Ve studii Padučeva 1985 (předchozí varianty Padučeva1979) je na základě logiko-sémantických kritérií propracován systém tzv. denotačních statusů[1] (podle Arutjunové je to typ reference) jmenné fráze ve všech možných pozicích.

Souvislost mezi jazykovým významem a referencí je různá u různých typů výrazů schopných reference. Na základě toho kriteria Padučeva dělí jména na čtyři skupiny: (1) vlastní jména, u kterých reference je založena na pragmatických znalostech mluvčího o světě, (2) deiktické výrazy, které nabývají (vždy stejný) význam v konkrétním mluvním aktu, (3) deskripce (kombinace obecného jména a deiktických výrazů) a (4) obecná jména, která nemají vlastní referenci a nabývají ji jenom v kombinací s výrazy deiktickými.

Jmenná fráze se podle Padučevové skládá z obecného jména (общее имя) a aktualizátora (актуализатор). Obecné jméno je slovníková jednotka, která v případě předmětného jména má extensionál (množinu objektů, které může označovat daná NP), a není zapojena v čase a místě. Aktualizátory jsou jazykové jednotky (slova nebo komponenty věty), které dělají z obecného jména aktualizovanou NP, zapojenou do kontextu a obsahující časoprostorové charakteristiky. Jako aktualizátory mohou sloužit např. různé typy zájmen (*ten, takový, každý, nějaký* apod.). Aktualizátor může mít nulovou hodnotu, tj. nemusí být ve větě explicitně vyjádřen (*Lékař přišel až večer*[2].), na druhé straně však aktualizovaná NP se může skládat jenom s aktualizátoru bez obecného jména, jak je to v případě substantivních použití zájmen (např. substantivní *to*, *já* apod.) nebo aktualizátor se napojí na již aktualizovanou NP (*některý z těch studentů*). Denotační status NP se určuje především významem aktualizátora dané NP, přičemž u různých druhů NP se jejich denotační status určuje s různou mírou jednoznačnosti (anglický král – vždy závisí na kontextu, vlastní jména – jediná interpretace) (srov. Padučeva2001, 84).

Klasifikace denotačních statusů podle Padučevové odděluje především tzv. termové (substantivní, věcné, předmětné) použití NP od použití predikativního.

Při predikativním použití NP se neprovádí reference na objekt, avšak jinému objektu se připisuje vlastnost. Stává se to tak v případech, kdy NP je součástí přísudku nebo se nachází v apozici. Srov. např. NP *lékař* ve větě *Иван врач*. „*Ivan je lékař*“ a NP *krasavice* ve větě *У него была дочь красавица* – dosl. «*Měl dceru krasavici*». Zvlášť se vyčleňuje skupina tzv. autonymních použití, kde NP má kleslý referent, jako např. ve větě *Муж просто звал ее Наташей.* „*Manžel ji říkal prostě Nataša*“. Za predikativní se považují také NP typu *реки, которые зимой замерзают* ve větě *Есть реки, которые зимой замерзают*, které by logicky spíše mohly mít existenciální status (viz dále).

Substantivní NP se dále rozdělují na jména se specifickou referencí (tzv. konkrétně-referenční, singulativní) a nereferenční jména.

Jména se specifickou referencí individualizují objekt, např. ve větě *Окно было маленькое и узкое* «*Okno bylo malé a úzké*» referent NP je již „vybrán“ z množiny všech významů lexémů *okno* a představuje určité okno v určitém místě. Uvnitř třídy jmenných skupin se singulativní referencí se NP dále klasifikují na základě rysu „± určenost“ (určenost objektu zároveň pro mluvčího a adresáta) a „± slabá určenost“ (určenost jenom z hlediska mluvčího).

Silná určenost NP souvisí s presumpcí existence a jedinečností objektu ve společném kontextu mluvčího a adresáta, tj. adresát vždy předpokládá, že adresát to bude takto interpretovat. Jedinečnost přitom může vyplývat z významu obecného jména (Srov. Лучшая из моих картин находится в Лувре „Můj nejlepší obraz je vystaven v Louvru“), je součástí aktualizátora (pokud se

[1] denotační status – typ referenčního zaměření NP, Padučeva převzala od Geach1962 (Padučeva2001, 83)

[2] pokud není uvedeno jinak, příklady X-X pochází z Padučeva2001

na objekt přímo ukazuje, jako např. na knihu ve větě *Я прочел эту книгу* "Přečetl jsem tuto knihu" nebo ve větě Ty книгу, которую ты мне дал, я уже прочел, kde je presumce ‚existuje jediná kniha, kterou jsi mi dal') nebo jsou jiné příčiny pro jednoznačnou identifikaci objektu, jako např. existence objektu ve společném *поле зрения???* mluvčího a adresáta apod. (o tom Padučeva zmiňuje, ale nerozpisuje to podrobně.). Určitost může být mimotextová a textová, přičemž pro textovou určenost není podmínkou bezprostřední použití dané NP v předchozím kontextu, ale může vzniknout i situací, kterou text generuje, srov. určenost NP *дорога* „silnice" v kontextu Он возвращался домой поздно. Дорога была плохо освещена. „Vrácel se domů pozdě. Silnice byla špatně osvícena".

Neurčité NP jsou v klasifikaci Padučevové dále rozděleny podle rysu „± slabá určenost". Slabou určenost ukazuje Padučeva na příkladech typu Он хочет жениться на одной иностранке. „Chce se oženit s jednou cizinkou", kde adresát pravděpodobně nebude vědět, o kterou cizinku jde, zatímco pro mluvčího to je známý objekt. NP s rysem „– slabá určenost" jsou neurčité také pro mluvčího, srov. např. *Иван читает какой-то учебник „Ivan čte nějakou učebnici"*. Rys „± slabá určenost" bývá často neutralizován, kdy z kontextu nepoznáme, je-li daná NP je pro mluvčího určitá, srov. Иван подрался с милиционером. „Ivan se popral s policajtem". (*odkaz na to, kde o tom bude ještě něco psát*).

Nereferenční NP nereferují na vybraný mimojazykový objekty a podle Padučevové podrobnější analýza nereferenčních NP umožňuje jejich další klasifikaci. V rámci nereferenčních NP se vyčleňují atributivní, univerzální, existenciální a generické NP.

Atributivní NP lze ilustrovat NP *Убийца Смита* ve větě *Убийца Смита сумасшедший* „*Vrah Smitha je blázen*" v takové interpretaci, když mluvčí má presumpci existence a jedinečnosti vraha, ale nemíní žádného konkrétního člověka (něco jako 'ten, kdo zavraždil Smitha, je blázen'). Srov. také některé NP ve větách obecné povahy jako *Самый сильный человек в мире не в состоянии поднять больше 200 кг*. «*Ani nejsilnější člověk na světě nedokáže zvednout víc, než 200 kg.*» *Тот, кто победит в этой борьбе, не избежит нечестных приемов* «*Ten, kdo vyhraje ten boj, nevyhne se ???*». apod.

Univerzální NP mají ve významu velký kvantifikátor, tj. označují všechny objekty abstraktní třídy, která představuje extensionál daného jména, např. *Все дети любят мороженое* «*Všechny děti mají rády zmrzlinu*».

Existenciální NP se používají v situaci, když se mluví o objektu, který patří do množiny objektů podobného typu a přitom není individualizován, tj. není vybrán z dané třídy. Padučeva vyčleňuje tři typy existenciální reference: distributivní, nekonkrétní a obecněexistenciální.

Distributivní NP (např. *Иногда кто-нибудь из нас его навещает* «*Občas ho někdo z nás navštíví*») referují k účastníkům, které jsou rozděleni na určitou množinu stejnorodých situací.

Nekonkrétní NP jsou takové NP, které vystupují v kontextech bez afirmativity. Padučeva uvádí seznamy typu kontextu, které vytvářejí podklad pro takové použití. Jsou to např. modální slova typu *může, chce, musí* aj.; rozkazovací způsob, budoucí čas, otázka, negace (včetně negace uvnitř lexému: *odmítat, nezbytně, zakázat*), disjunkce, podmínka, cíl, nejistota, předpoklad; některé propoziční predikáty: *chtít, myslet* aj. a performativní slovesa (*prosím, slibuji*). Srov. např. *Джон хочет жениться на какой-нибудь иностранке* «*Chce se oženit s nějakou cizinkou*», v případě, že se ještě neseznámili; *Он ищет новую секретаршу* «*Hledá novou sakretařku*» apod.

Obecně-existenciální NP se vyskytují v textu, když se mluví o objektech s určitými společnými vlastnostmi a referují např. k neurčitému počtu objektů určité třídy, aniž by ty objekty byly individualizovány. Srov. *Некоторые вещи портятся при перевозке*. «*Některé věcí se mohou poškodit při stěhování*»

Do nereferenčních NP podle klasifikace Padučevové (1985) patří také generické NP.

Krátce systém denotačních statusů Padučevové lze znázornit v následující tabulce.

substantivní	referenční	+určitost	*Я прочел эту книгу "Přečetl jsem tuto knihu"*	
		- určitost	+ slabá určitost	*Он хочет жениться на одной иностранке.*

				„Chce se oženit s jednou cizinkou“
			- slabá určitost	*Иван читает какой-то учебник „Ivan čte nějakou učebnici“*
	nereferenční	atributivní	*Убийца Смита сумасшедший „Vrah Smitha je blázen*“	
		univerzální	*Все дети любят мороженое* «*Všem dětem chutná zmrzlina*».	
		existenciální	distributivní	*Иногда кто-нибудь из нас его навещает* «*Občas ho někdo z nás navštíví*»
			nekonkrétní	*Он ищет новую секретаршу* «*Hledá novou sakretařku*»
			obecně-existenciální	*Некоторые вещи портятся при перевозке*. «*Některé věcí se mohou poškodit při stěhování*»
		generické	*Скорпион похож на кузнечика «Štír vypadá jako koník»*	
predikativní	*Иван врач*. „*Ivan je lékař*“			

Problém s klasifikací Padučevové je však v tom, že všechny její příklady jsou konkrétní předmětná jména, která sice krásně znázorňují typy, ale zanechávají spoustu otázek u NP s méně konkrétním významem. Kromě toho v ukázkách typů reference nejsou anaforické páry, ale jednotlivé výpovědi. Proto třeba není jasné, kam se podle Padučevové zařadí anaforická NP s antecedentem s nespecifickou referencí. Z ustní diskuze a její distribuce určenosti jsem pochopila, že pokud má daná NP určitý identifikátor, nemůže být již pojatá jako nespecifická (určenost je podle Padučevové příznak, který je aktuální jenom pro NP se specifickou referencí). Při anaforickém opakování se tedy postuluje svět, ve kterém se daná NP už chápe jako specifická.[3]

X.X.X.X. Šmelev1996

V popisu mechanismů reference Šmelev vychází z dvou základních pojmů – denotativní prostor (денотативное пространство) a relevantní denotativní prostor (релевантное денотативное пространство). Denotativní prostor je jakýkoli úsek mimojazykové skutečnosti. Pro kterýkoli jazykový výraz, který se používá v promluvě, relevantní je ten denotativní prostor, ve kterém je určen referent daného jazykového výrazu. (Šmelev 1996,23). Například relevantním denotačním prostorem pro NP *ректор* ve větě *Я расскажу об этом ректору* je v případě přirozené interpretace denotační prostor univerzity, ve které se nachází účastníci komunikace.
Srovnat s Bühler (u něj se to také nějak jmenuje).

Rozbor referenčních mechanizmů v Šmelev je proveden v rámcích tzv. neokauzální teorie reference, která se zakládá na postupném obohacování „мысленного досье“ adresáta na daný referent. Ukazuje to na příkladě vlastních jmen, kde na začátku konverzace adresát nemusí o referentu, označeném daným jménem, vědět nic, potom však postupně dostává další informace, čímž se jeho „мысленное досье“ doplňuje. Tedy provádí se kauzální řetězec od introduktivní promluvy (typu *To je Karel*) dále do dalších a dalších informací o adresátovi.(Šmelev 1996,33).

Mimojazykové objekty se třídí podle Šmeleva na třídy („классы“) a prvky („индивиды”). Třída je otevřená nepočetná množina objektů, prvky jsou jednotlivé objekty nebo uzavřené množiny objektů. Podle toho se typy reference dělí na generickou („генерализованную“) a individuální („индивидную“). Zvlášť se vyčleňuje abstraktně-individuální referenci. Referent takové NP je sice individuální ale je použit ve výpovědí, která nemá časoprostorové charakteristiky (např. NP *собака* ve větě *Собака любит Ивана* nebo v pozici objektu *пойти в магазин, лечь в больницу*). (Šmelev 43sl.)

Uvnitř generické reference se dále vyčleňují obecněgenerické NP („общеродовые“) a obecněexistenciální NP („общеэкзистенциальные“). Obecněgenerické odrazují k celé třídě objektů (*Собака – друг человека; Все дети*[4] *любят сказки*). Obecněexistenciální NP odkazují k některé

[3]Je to ale jenom můj pohled na její názor, sama o tom v Padučeva1985 nepíše.
[4] V klasifikaci Padučevové by daná NP měla univerzální referenční charakteristiku

(možná také neomezené) částí dané třídy (např. *Некоторые логики разбираются в лингвистике*). Důležitá pro náš výzkum je poznámka Šmeleva, že **při anaforickém opakování daná NP bude mít již obecněgenerickou referenci**, jako např. NP *Эти логики* v následující promluvě: *Некоторые логики разбираются в лингвистике. Эти логики обладают хорошим языковым чутьем.*

Generické, abstraktně-individuální a individuální jmenné fráze všechny mohou mít určitou a neurčitou referenci. Tím se koncepce Šmeleva liší od teorie reference Padučevové, která o kategirii určenosti-neurčenosti mluví jenom v rámci konkrétní reference (v terminologii Šmeleva - individuální).Takový přístup k určenosti se zdá být výhodnější, protože se i jmenné fráze s generickou referenční interpretací dostávají např. do anaforické pozice, používají se s identifikátory určeností a chovají se velmi podobně NP s konkrétní referencí. (Srov. – *dát nějaký příklad, rozepsat elegantněji*)

Šmelev 1996 v rámci své neokauzální teorie reference zavádí ještě jeden velmi důležitý pro náš další výzkum pojem indexní reference (индексальная референция). V případě indexní reference typ reference je určen typem zájmena, které se při té NP vždy používá, přičemž deskriptivní význam NP vůbec nemá (jako např. osobní zájmena) nebo ten význam není relevantní pro identifikaci daného referenta (jako např. v případech, kde deskriptivní význam nestačí pro identifikaci – *Какой-то дурак все испортил, Дай-ка мне эту штуку*)

X.X.X.X. B. Palek1988

Podobně jako Padučeva1985 NP chápe jako spojení komponentu nominálního, který představuje obecný výraz, a komponentu instauračního (v terminologii Padučevové - aktualizátor), který stanoví, že jmenná fráze označuje nějaký denotát, tj. realizuje referenci daného nominálního komponentu. Prostředky vyjádření denotace jsou v přirozeném jazyce výrazy, které Palek nazývá instaurátory. Pojem „instaurátor“ je podobný aktualizátoru ve smyslu Padučevové, je však o něco dále rozvit tím, že jako morfematický aktualizátor vystupuje např. osobní koncovky sloves a číslo substantiv, syntaktická pozice výrazu ve větě (např. subjektová) aj.

Cílem práce Palka je pak provést zevrubnou klasifikaci instaurátoru pro češtinu. Zatímco Padučeva rozebírá funkce celých NP v různých referenčních kontextech, Palek provádí analýzu vysloveně instaurátorů samotných na základě následujících kriterií: autosémantičnost (instaurátory, které samy referují na denotát)/synsémantičnost (spolu s jinými výrazy podílejí na určení příslušné instaurační funkce), jednoduchost (skládá se z jednoho výrazu – *tam, takový* apod.)/složenost (skládají se z více výrazů – *takový ten, poslední z nich*), závislost/nezávislost, samostatnost/nesamostatnost apod.

Instaurátor je závislý, pokud jeho výběr pro danou NP je ovlivněn přítomností jiného instaurátoru v jiné NP. Mezi nezávislé instaurátory patří např. kvantifikátory *každý, některý* apod. Závislé instaurátory se člení do dvou skupin – identifikátory (potvrzují již zmíněný denotát) a alternátory (vymezují jiné denotáty než ty, které byly zmíněny). Závislost realizovaná prostřednictvím závislých instaurátorů umožňuje hledat v textu anaforické a alternační posloupnosti. Samostatné instaurátory vystupují ve větě bez obecného výrazu (jako např. *já, tam* apod.) nesamostatné instaurátory rozvijí řídící výraz, v češtině jsou to především zájmena s adjektivním skloňováním (*každý, nějaký* apod.).

Palkova analýza referenčních principů je zaměřena především na popis fungování anaforických vztahů v textů. Při stanovení vztahů mezi denotáty v anaforických řetězcích Palek zmiňuje identitu, inkluzi (referent antecedenta v sobě zahrnuje referent postcedenta), členství a rozdíl (pak se používají alternátory) avšak podrobnější klasifikace se v práci neprovádí (*ověřit – ale zatím jsem to tam nenašla*). Co se týče přímé analýzy možných typů reference, ta se neprovádí systematicky. Rozlišuje se mezi denotáty-konstanty (pokud situace nebo jev jsou časoprostorově určeny) a proměnnými („situace, která je pouze součástí intence mluvčího“) avšak vzápětí se tvrdí, že

„konstatnost/proměnnost subjektu/objektu se vzájemně neovlivňují, nekladou žádná sémantická omezení na věty a týkají se pouze sémantické interpretace anaforických vztahů“ [5].

Následuje rozbor příkladů z oblasti anafory – jak uvnitř věty a souvětí, tak i uvnitř textu. Vysvětlují se různé kontexty, kde jde vynechat Subjekt, a bude se vyrozumívat z kontextu, kde může být on nebo ten. Dělá se to na jednotlivých příkladech, vypadá jako předpoklady, není ověřeno na korpusu. (s.75n.)

Pro analýzu referenční struktury textu jsou navrženy dva principy: analytický a syntetický. V rámci principu analytického se pomocí složitého systému založeného na pojmech matematické logiky a relací identity (eq) - různosti (neg), inkluze (inc) – neinkluze (ninc), disjunkce (dis) – nedisjunkce (ndis) a členství (meb) – nečlenství (nmeb) provádí analýza vztahů mezi denotačními frázemi. Výsledkem jsou anaforické textové vzorce a pokus o vymezení pojmu text. Syntetický model představuje generování textové jednotky a možných referenčních vztahů v ní. Oba modely se navzájem doplňují a předpokládají.

co vyplývá z Palkovy klasifikace instaurátorů?

Z Palkovy klasifikace instaurátorů vyplývá, že vztahy exoforické, endoforické (katafora a anafora), alternace a negace jsou vymezeny jenom pro referenční pozice, tj. neurčují se pro vztahy predikační povahy a pro apozici (i když apozici jako nereferenční pozici explicitně nezmiňuje). Avšak jak se ukazuje z příkladů, uvedených v dalším výkladu (s.59n.), NP v opozici může vystupovat jako antecedent pro anaforu. Srov. (*přepsat jeho příklad*) Předpokládám však, že to není typycké.

Další podstatná pro naši analýzu informace, která vyplývá z klasifikace, je že pro postcedent v exoforickém vztahu je obligatorní přítomnost identifikátoru – samostatného (v podobě substantivní nebo adverbiální) nebo nesamostatného v podobě adjektivní. Předpokládám, že se s tím dá docela diskutovat – *napsat pár protipříkladů, kde se žádný identifikátor nepoužívá, a přesto exoforický odkaz tam nacházíme, který se vyrozumívá z kontextu*. Jak dále vyplývá z tabulky (s.48) a popisu, exoforický odkaz se liší o endoforického jenom tím, že jeho antecedent není přímo zmíněn v předcházejícím textu, ale vyrozumívá se ze situace, zatímco při vztahu endoforickém, antecedent a postcedent „jsou v textu explicitně vyjádřeny a jim odpovídajícím denotačním vztahem je identita nebo incidence denotátů“ (s.47). Zajímavá jsou i omezení pro endoforický typ vztahu. Např. pro kataforu se uvádí, že antecedent při kataforickém vztahu může mít u sebe jenom samostatný instaurátor, což samozřejmě záleží na chápání toho, co je katafora, ale při dost širokém chápání, to tak asi nebude. (zajímavé je, že při dalším výkladu uvádí příklad na kataforu, kde identifikátor zrovna není samostatný - *...přírodní památky mají ten zvyk, že existují většinou v končinách odlehlých* (s.113)) Srov. např. něco jako (*uvést příklad*) *Viděl takové zvíře, které ještě nikdo neviděl*. Co se týče anaforického vztahu, postcedent může mít jak samostatný tak i nesamostatný identifikátor.

Nedostatkem Palkova přístupu je to, že při rozboru vnitrotextových referenčních vztahů se orientuje skoro výjimečně jenom na anaforu a na analýzu vztahů mezi antecedenty a postcedentem. Zdá se však, že se tím ztrácí velké procento textových vztahů, které za anaforu označit nelze. Na našich příkladech je to dost dobře vidět. (*uvést nějakou statistiku*). K referenci NP jako takové se přihlíží jen velice málo, ta je rovnou spojená s anaforou, přičemž za anaforu se považují jenom případy, kdy postcedent obsahuje příslušný identifikátor.
nejde o referenci konkrétní NP v textů

X.X.X.X. Yokoyma

……. (dodělat) Teorie průniků množin Yokoyamy je v podstatě přesnější představa denotačních prostorů Šmeleva. To se dá názorně demonstrovat na příkladě reference vlastních jmen. (*popsat oba ty výklady*)

[5] Palek1988,78

X.X.X.X. Mendoza2004

Habilitační práce Imke Mendoza 2004 si neklade za cíl nově zpracovat teorii reference, ale popsat systém jmenné determinace a jejich prostředků v současné polštině. Avšak k tomu je zapotřebí mít přesné referenční nástroje. Proto autorka popisuje a srovnává některé existující teorie a nakonec vytváří vlastní přehled, který se mi zdá velice přesný a zajímavý. Stručně ho tedy představíme. Denotativní status jmenné fráze se podle Mendoza2004 skladá z jejích referenčních a textově pragmatických vlastností. (rozepsat - viz Mendoza s.69) Autorka tvrdí, že tyto dvě roviny jsou mezi sebou neoddělitelně propojené a textová určenost je často závislá na typu reference dané NP, nelze je tedy zkoumat zvlášť.

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

- distributivní a kolektivní, kde se odkazuje k celé počitatelné a uzavřené množině objektů ($\forall$ (y) $\exists$(x) G(x,y) a $\exists$ (x) $\forall$ (y) G(x,y): *Die Elefanten* sterben aus.),
- reference na otevřené množiny neby třídy (Referenz auf Klassen), která se dále člení na univerzální ($\forall$x (K(x) $\rightarrow$ S(x)): *Alle Kinder* essen gerne Schokolade.), generickou (Das Auto ist *des Deutschen* liebstes Kind.) a existenciální typy reference. Generická reference se pokladá za typ zvláště komplikovaný a nejednotný, proto autorka pokládá za smysluplné rozdělit ho ne několik dalších podtypů. Kromě toho různé podtypy se mohou v jazyce vyjadřovat různými způsoby. Tyto podtypy jsou: generická reference na typ (typen-generische), tj. odkaz na prototypického představitele dané třídy; generická reference na trídu (klassen-generische), tj. odkaz na libovolný element dané třídy reprezentativní generická reference (repräsentativ-generische), která vystupuje v nezobecňujících výpovědích.
- reference na prvek, která může být specifická (NP má konkrétní denotát v popisovaném světě) a nespecifická (NP takový denotát nemá). Kontexty pro nespecifickou referenci jsou např. nereální kontexty, otázky, kondicionál, negace apod. Zvlášť autorka upozorňuje na rozdíl mezi referencí na objekt a tzv. referencí na funkci (Referenz auf Rollen), která v některých kontextech může být i syntakticky relevantní. Za referenci na funkci autorka považuje takové užití NP, které se tradičně (Padučeva, Donnellan) považuje za tzv. atributivní určité deskripce, srov. NP *обидчик* ve větě *On nakazal obidčika*. (Padučeva 1985, cit. podle Mendoza2004)

Řešení

Při anotaci rozšířené koreference na PDT 2.0 rozlišujeme referenční a nereferenční NP podle Mendoza 2004, pričemž takovým způsobem, že nereferenční NP neanotujeme. Za nereferenční tedy považujeme a neanotujeme:

- koreferenční vztah mezi subjektem a predikátovou částí výpovědi (*Petr je programátor*)
- koreferenční vztah mezi členy apozice (*Petr, náš programátor, má zítra státnice*)
- koreferenci na uzel s funktorem ID, protože NP v takových případech nereferuje na objekt vnějšího světa, ale sama na sebe – autonymní použití. Srov.

> (4/In95047_061) Podle těchto zpráv nějaká firma na naše území umísťuje německou delikventní mládež , která zde páchá kriminální činy a ohrožuje starousedlíky .
> (15/In95047_061) V Košťanech totiž zakoupila dům firma (coref_text, typ=0 na „firma“) Struktura (funktor ID), která se u nás rozmísťováním německých chlapců zabývá .

Takové řešení nám umožňuje dostatečně bezproblémové rozdělení NP na referenční a nereferenční (ještě to však nemám úplně vyřešeno s ID), ale ochuzuje nás o rozlišení identifikačních a predikačních konstrukcí, když analyzovaná NP je ve větě v pozici přísudku. Ukazuje se, že rozhodnutí mezi dvěma významy predikace (přisuzovaní kvality) a identifikace referenta subjektu a jmenné částí přísudku není vždy jednoznačné a vyžaduje hlubší analýzu.(Můžu k tomu dodat přehled literatury, ale anotovat bych to přesto nechtěla. Někdo z počit. lingvistů – uvést kdo – to ale dělá)

Další aktuální otázka z teorie reference, kterou musíme vyřešit pro anotaci rozšířené koreference je rozlišování specifické a nespecifické reference, zvlášť u NP, které vystupují jako koreferující člen s antecedentem s nespecifickou referencí. NP má v textu specifickou referenci, pokud objekt je vybrán, individualizován, v reálném nebo fiktivním světě, který je vytvořen daným textem. Pokud objekt není vybrán, jeho reference je nespecifická, i kdyby vystupoval jako koreferující člen s určitým identifikátorem. Srov. v následujícím příkladě NP „tento podnikatel“ má nespecifickou referenci, protože objekt „podnikatel“ není vybrán ze řady jiných podnikatelů.

> (19/In94208_11) Tímto faktorem je podnikatel - inovátor , který se snaží o zisk , a proto logicky nemůže existovat ve stavu statiky , která nezná ani zisk , ani ztrátu .
> (20/In94208_11) Tento podnikatel {coref_text, typ=NR na „podnikatel“} se od manažera liší tím , že zavádí nové kombinace výrobních faktorů , kdežto manažer je jen rutinně kombinuje na bázi dané techniky .

Platí dokonce pravidlo, že pokud anaforická generická NP je hyperonym ve vztahu k antecedentu, použití aktualizátora pro zachování koreference s antecedentem je nezbytné. Srov.

> *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, začínal rojit*. <doc S|NOV|1994|borivoj> (příklad ze SYN2000)

Je to poněkud problematická záležitost. Jako argument pro to, že reference uvedené NP je NR je také např. to, že generické NP běžně vystupují jako antecedenty v anaforickém vztahu. Srov. něco jako *Děti často dělají blbosti. Asi také proto, že nevědí, že to jsou blbosti*. Jinak to bude u NP s nespecifickou referencí, které nejsou generické, třeba u NP v neoznamovacích kontextech (jako *Němec* ve větě *Chtěla bys si vzít Němce?*). Pro takové NP anaforická pravidla fungují pouze v dosahu identifikátoru, tedy přez tečku v jiné výpovědi už fungovat nebudou. Z toho plyne, že by se nám v naších textech neměly objevit NP s identifikátorem určenoti s nespecifickou negenerickou referencí v pozici koreferujícího členu. Pokud se něco takového stane, pak už to bude NP se specifickou referenci (objekt už bude vybrán). Srov. v anaforické pozici *A že tam někde nahoře musí být život, stejně jako tady, a ať už ty bytosti vypadají jakkoliv, dívají se na nás.* <doc S|NOV|1996|matskol>, nebo *Přesto si značky mohl všimnout jen někdo velice pozorný […] a ani ten velice pozorný člověk by jim patrně nepřikládal žádný význam.* Může se však stát, že se v pozici koreferujícího členu objeví třeba tatéž neurčitá NP se stejným typem nespecifické reference, a pak ji označíme jako coref_text, typ=NR. Nebude to ovšem anaforický vztah, ale pouhá coreference. *příklad* Možná by bylo logičtěji takové vztahy neoznačovat – koreference tam v uzkém smyslu není (jakápak koreference, když obě NP referují na nevybraný objekt), anaforický vztah přesahující větu je vůbec vyloučen, tak k čemu nám vůbec je? Ale trochu se obávám, že odlišit v reálném textu nereferenční použití generické a negenerické je poněkud komplikovaný a časově náročný úkol.

X.X.X. Predikace vs. Identifikace

S problémem rozlišení identifikačních a predikačních konstrukcí setkáváme v případě, když analyzovaná NP je ve větě v pozici přísudku. Ukazuje se, že rozhodnutí mezi dvěma významy predikace (přisuzovaní kvality) a identifikace referenta subjektu a jmenné částí přísudku není vždy jednoznačné a vyžaduje hlubší analýzu. V následující kapitole se pokusíme na základě existující literatury k danému tématu najít teoretické vymezení těchto dvou typů konstrukcí.

Názory badatelů na hranici mezi predikací a identifikací se značně rozcházejí. Existuje širší a užší pojetí identifikace.

Nejširší pojetí identifikaci najdeme v **Lavric 2001**, 63ff. (cit. podle Mendoza), která za predikaci považuje jenom takové konstrukce, ve kterých není možné užit členu, jako např. NP *Determinantensemantikerin* ve větě *Ich bin Determinantensemantikerin*. Přísudková substantiva s neurčitým členem jsou považovány za referenční s neurčitou referenci. Jako identifikační konstrukce se hodnotí dokonce věty s generickou jmennou frází v přísudku typu *ein Insekt* ve větě *Die Heuschrecke ist ein Insekt*. V tom případě identifikace se provádí mezi třídami objektů. Toto široké pojetí identifikace kritizuje (Mendoza2004, 73n.). Namítá, že v případě zařazení vět s neurčitým členem v přísudku mezi identifikační struktury nastává situace, když jako predikace se interpretují věty typu *Marek ist Lehrer* a *Johann ist gut* a jako identifikace s neurčitým objektem věty typu *Johann ist ein guter Mensch*. a *Marek ist ein guter Lehrer*. Z toho, že německá věta bez členu **Marek ist guter Lehrer* je agramatická, plyne, že není možné vyjádřit predikační význam pomocí NP, která obsahuje přídavné jméno, což není úplně logické. Kromě toho, orientace definice predikačních NP na existenci ve větě členu je příliš zaměřena na konkrétní jazyk. V angličtině situace bude dost odlišná, srov. agramatičnost angl. **He is teacher*.

Jiný názor nacházíme v prácích lingvistů moskevské sémantické školy – Paducheva1987, Šmelev1996, Arutjunova1976 aj. V těchto pracích se identifikace („высказывание идентификации“) chápe také dost široce, ale rozdíl mezi identifikací a predikací se provádí na úrovní sémantické za pomoci některých lexikálně syntaktických kritérií.

Arutjunova1976 analyzuje sémantickou a referenční strukturu relace identifikace. Vyčleňuje se několik typů situací identifikace, např. „situace detektivního hledání“ (*убийца старухи есть Раскольников*), situace uskutečnění snu (*Это как раз и есть то, что нам нужно; Вы и есть нужный человек*), situace „poznávání“(*Иван Иваныч! да ведь это ты! ты! ты!*) a několik dalších. Situace identifikace se vyjadřuje v IE jazycích pomocí tzv. identifikačních konstrukcí („предложения тождества“). V rámcích identifikačních vět se rozlišují nominativní (kódová) totožnost, v případě když se deklaruje stejná schopnost k referenci dvou rozlišných nominací (Цицерон есть Туллий *dodat překlad*) a tzv. denotativní totožnost, když se deklaruje totožnost objektu sebe (Бьюсь об заклад, если это не тот самый сорванец, который увязался за нами на мосту *dodat překlad*). Od obou komponentů identifikační konstrukce se vyžaduje konkrétní reference (Arutjunova1976, 292).

Za identifikační se považují všechny výpovědi, které mají vlastní jméno v rématu, protože mohou být transformovány na konstrukci typu N1– je – N1. Srov. např. Во время катастрофы пострадал Иванов → Пострадавший - Иванов *dodat překlad*.

Testy na klasifikací/identifikací podle Arutjunova1976, 292

(1) možnost použit určitého determinátoru s oběma komponenty identifikační konstrukce, zatímco v klasifikačních (podle Arutjunové tzv. „inkluzívních“) větách druhý komponent buď nemá žádný identifikátor nebo má neurčitý člen. Srov. angl. *Peter is a writer* (predikace) vs. *Peter is the author of this novel*. (identifikace)

(2) v identifikačních větách oba komponenty mohou být zastoupeny zájmenem, srov. např. Этот молодой корнет и есть девица Дурова → Он и есть она. (Arutjunova1976, 311);

(3) test pokračování textu: pokud zkoumaná konstrukce je klasifikační, její druhý komponent nemůže sloužit subjektem následující výpovědí, protože nemá vlastní referenci, zatímco v identifikační výpovědi to je možné. Srov. např. identifikační větu Мой учитель – Джонс. pro kterou pokračování typu *Мой учитель (Джонс) преподает мне математику* zní celkem přirozeně s oběma podměty. Avšak pokud v první větě změníme pořádek členu na Джонс – мой учитель, dostaneme klasifikační větu a subjekt pokračování Мой учитель преподает мне математику nebude koreferenční se subjektem první věty.

(4) test negace – pokud na identifikační větu naložíme negaci, zápor se bude vztahovat jenom na tvrzení o tom, že referenty obou komponentu věty jsou identické. Presupozice jejich existence zůstane nezměněná. Srov. např. větou Ложно, что в Бородинском сражении победителем был

Наполеон *dodat překlad* (Negace naložena na identifikační větu Победитель Бородинского сражения – Наполеон *dodat překlad*) se nepopírá existence ani Napoleona ani vítěze Borodinského boje, ale jenom totožnost referentů těchto dvou výrazů.

(5) vratnost (reverzibilita?) komponentů identifikačních konstrukcí typu Venuše je Jitrnice ↔ Jitrnice je Venuše.

Pokus najít formální rozdíl oproti charakterizačním větám v nemožnosti použití v identifikačních konstrukcích přísudkového Instrumentálu se však autorce nepodařil. Jsou totiž správné i identifikace s Instr., srov. Этим клоуном был мой сослуживец (*dodat překlad*) vedle Этот клоун был мой сослуживец. (*dodat překlad*) Dokonce se mezi těmito větami nachází jistý sémantický rozdíl. Podle Arutjunové je možnost instrumentálového přísudku však sekundární a je podmíněna kontaminací slovesného rámce slovesa *оказаться* (*dodat překlad*) (Arutjunova1976, 325).

Dopsat pak něco, jak ty testy fungují na našem materiálu, třeba něco nefunguje. Určitě bude problém hned s tím prvním na determinátory.

Padučeva1987 v popisu identifikačních vět («предложения тождества») vychází z identičnosti referentů prvního a druhého komponentu konstrukce. Důležité přitom je, aby oba komponenty měly specifickou (konkrétní) referenci[6]. Teoreticky by bylo možné za identifikaci považovat i konstrukce, ve kterých vystupují generické NP (srov. např. *В маленьком населенном пункте главный праздник - ярмарка*), ale právě tím, že dané NP nemají konkrétně referenční status, se takové NP z klasifikace vylučují. (Srov. Padučeva1987, s.154, ale také Weiss1978,245). Za identifikační se dále nepovažují následující typy vět:

1. různá metajazyková užití – např. vysvětlení obecných pojmů (Аксиома – это истина, не требуюшая доказательств *dodat překlad*), vnitrojazykové překlady (Октаэдр – это восьмигранник *dodat překlad*) apod.
2. taxonomická identifikace – např. Первый урок была история *dodat překlad*; Президент Филиппин – женщина (*tj. někdo, kdo je žena*) *dodat překlad* apod. Za pravé identifikační konstrukce se považuje jenom identifikace substanční.
3. identita s abstraktními pojmy (spravedlnost, láska), protože se u nich špatně určuje referent. Události a procesy mohou být v identifikačním vztahu, ale nestává se to příliš často.
4. věty vyjadřující metaforickou identitu (*Государство – это я*; *Já jsem ta tma* (Šárka - poslední řádek v básničce ve vlaku))

Z hlediska aktuálního členění Padučeva upozorňuje na zdánlivou subjektivnost slovosledu v identifikačních větách. Říkáme-li větu „Утренняя звезда – это Венера“ (Jitrnice je Venuše), o Venuši se předpokládá, že adresát je ji schopen identifikovat, zatímco pro Jitrnici je zapotřebí další vysvětlení. Tím se však jedině znázorňuje fakt, že k tomu, aby bylo něco v tématu stačí kontextová zapojenost (viz dále o AČV *dodat odkaz*) a že to nemusí nutně souviset s aktuálními znalostmi adresáta o světě.

Na sémantické úrovni se podle Padučevové dá mluvit o čtyřech typech sémantické identifikace:

a) 1. komponent je atributivně použitá deskripce: Столица Перу – Лима;
b) provádí se shoda mezi lokálně a temporálně identifikovaným objektem a jeho obecným názvem (Эта освещенная магистраль – улица Кропоткина *dodat překlad*); nebo deskriptivní znalost objektu se identifikuje se samým objektem nebo jeho manifestací (Маяковский – это я *dodat překlad*).
c) Oba komponenty jsou obecná jména a „globální názvy“, adresátovi se „nabízí možnost změnit model světa, ve které odpovídající jména označují dva různé objekty na takový, kde označují jeden stejný objekt“[7]

[6] Srov. klasifikaci denotačních statusů v Padučeva1985, a Arutjunova 1976.

[7] „говорящий предлагает слушающему заменить модель мира, в которой соответствующие имена обозначают два разных объекта, на такую, в которой они обозначают один и тот же объект, ср. *Утренняя звезда и Венера – это одно и то же небесное тело* “ (Padučeva1987,161)

d) «устанавливается принадлежность двух ипостасей или срезов одному и тому же объекту» (Хозяин гостиницы – оценщик в городском ломбарде *dodat překlad*)

V Padučeva1987 se naznačuje, že rozdíly mezi uvedenými se projevují také v syntaxi (jako např. různé možnosti atributivních a komunikativních modelů, možnosti ne/použit výrazu *eto* jako součást druhého komponentu konstrukce, rozlišná pravidla užití zájmen v prvním komponentu), v morfologii (aplikace na časový systém ruštiny) apod. Tyto rozdíly se však dále nerozebírají.

Добавить критику на примерах. Может быть...

Šmelev1996 tvrdí, že cíl identifikační výpovědi je poskytnout adresátovi výpovědi možnost přesně lokalizovat referent prvního komponentu[8]. Podle Šmelev1996 s.177 identifikační výpovědi mohou být homonymní s výpověďmi s predikativní NP v přísudku, např. výpověď *Ivan – moj drug* (Ivan je můj kamarád) lze pochopit na jedné straně jako charakteristiku Ivana, na druhé straně jako odpověď na otázku *Kdo je Ivan?*, což je podle Šmeleva případ „vysvětlující" identifikace. Tato homonymie lze vyřešit pomocí perifráze:

Ivan – moj drug → Ivan mnje drug: predikace

Ivan – moj drug → Ivan – eto moj drug: identifikace

Za ukazatele identifikace lze považovat UZ *eto* v pozici prvního komponentu, a užívané zároveň s ním determinátory *etot, odin* jako součást druhého komponentu konstrukce[9], srov. *Это один мой друг*. Naopak NP na funkci charakterizační ukazuje podle Šmeleva osobní zájmeno ve funkci subjektu (*Он мой друг*) a některé další specifické prostředky, které zdůrazňují predikativnost druhého komponentu. Tato kritéria však neplatí vždy. Např. ve větách s *eto* typu *Это необыкновенный ребенок, Это талант* apod., které Šmelev pokládá za zvláštní druh identifikace, kde „kvalifikace se tváří jako identifikace"[10], je intuitivně přirozeněji postulovat vztah charakterizační (Srov. k tomu také Mendoza2004, 75). Podobných příkladů se však v jazyce najde víc než dost, čímž se jednoznačnost souvislostí *eto* ve funkci subjektu s identifikační povahy odpovídající výpovědí výrazně zpochybňuje.

Weiss1978 poukazuje na časté smíšení pojmů *Identität* (identita, totožnost, úplná shoda), ve kterých jde o totožnost referentů prvního a druhého komponentů konstrukce a *Identifikation* (identifikace, ztotožnění), které může být přítomné i ve výpovědích s predikační NP v druhém komponentu. Daný problém nastává často i pro češtinu – identifikace se dá rozumět neterminologicky i jako zařazení objektů do skupiny, rozebíráme však jenom to, čemu Weiss říká *Identität*. Identifikační věty se podle Weisse charakterizují dvěma kriterií: koreferencí referentů obou komponentu konstrukce a zvláštní komunikativní strukturou, o které tvrdí:

> ... die beiden verglichenen Ausdrücke sind kommunikativ nicht gleichwertig, insorefn als der Referent des ersten vom Sprecher als unbekannt, derjenige des zweiten als bekannt vorausgesetzt wird. (Weiss1978,228)
>
> „oba výrazy konstrukce nejsou z komunikativního hlediska rovnocenné; zatímco referent prvního výrazu je představen mluvčím jako neznámý, referent druhého výrazu je představen jako známý" (*překlad můj - ověřit*)

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

Situace nominalizace (Benennung) podle Weisse není identifikační. V potenciálně dvojznačných větách (např. *Я Распутин «Já jsem Rasputín»*) identifikaci poznáme podle toho, že její druhý

[8] k obrácené komunikativní struktuře těchto konstrukcí viz dále – Weiss1978

[9] z toho tvrzení dost jednoznačně plyne, že u NP s predikačním statusem se neočekává existence UZ – k tomu srov. následující úvahy o češtině. Explicitně to tvrdí možná jen Adamec, podívat se ještě někde (Berger, velká monografie), třeba se najde, ale implicitně je to všude, že v predikativní pozici determinátor stát nemůže. Najít ještě u Palka.

[10] «Такое употребление означает, что объект (чаще всего – лицо) рассматривается как «персонификация» указанного качества. [...] По существу здесь характеризация маскируется под идентификацию» (Šmelev1996,178).

komponent bude prezentován posluchači jako známý, tj. takový, kterému může být jednoznačně přisouzen referent, zatímco v nominalizačních větách je situace opačná: téma může být známé a KZ, zatímco réma nikoliv. Srov. jeho příklad *Жил-был один король. Короля звали Вася*. – č. *Byl jednou jeden král. Ten se jmenoval Vasja*. (Weiss1978,227)

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

a) syntax – důležitá je pozice ve větě. pokud obecné jméno je v identifikační konstrukci na začátku věty v pozici podmětu, zachovává se identifikační význam. Jakmile se přemístí na pozici přísudku, tento význam se ztrácí. (Weiss1978,232)
b) syntax – predikační větu můžeme parafrázovat jako normální predikaci: Я был автором этой статьи → Я написал эту статью, čes. (*překlad můj*) Byl jsem autorem toho článku → Napsal jsem ten článek. (Weiss1978,233)
c) syntax – na rozdíl od Arutjunové tvrdí, že určitý determinátor s druhým komponentem není jednoznačný ukazatel identifikace. Např. věta ru. *Он – убийца старухи*, něm. *Er ist der Mörder der alten Frau*, č. něco jako *On je ten, který zabil tu starou paní* (*udělat něco s českým překladem*) může být pojata dvěma způsoby: a) patří k množině lidí, kteří zabili tu starou paní; pravděpodobně tato množina obsahuje jenom ten jeden prvek. Význam věty odpovídá '(on) zabil tu starou paní'; b) je identický tomu člověku, kterého adresát zná jako toho, kdo zabil tu starší paní. Význam věty neodpovídá '(on) zabil tu starou paní'. V prvním případě jde o predikaci, v druhém – o identifikaci. (Weiss1978,237)
d) dobré lexikální kritérium na identifikaci pro ruštinu, které funguje v případě opačného T-R pořadí – srov. k tomu Padučeva1987,157n., Arutjunova1976,312 – možnost dodat „и есть", jako např. ve větě *Зевс – это Юпитер*). Relace typu *Identität* je možná jenom pro referenční NP, nikoliv pro výpovědi typu *Кимарить – это спать, Гардероб – это то же, что платяной шкаф*, kde se postuluje vztah synonymie.

Podstatu identifikační výpovědi předvádí Weiss tabulky se syntaktickou a referenční informací o prvním a druhém komponentech konstrukce. Tuto tabulku v trochu upravené podobě uvádí Padučeva1987. Tady podávám variantu, která je více podobná té, kterou uvádí Padučeva, abych nemusela řešit větněčlenskou platnost komponentů identifikační konstrukce:

1. komponent identifikační konstrukce, X	2. komponent identifikační konstrukce, Y
téma	réma
kontextově zapojen	není kontextově zapojen
není (dostatečně) určitý	známý, určitý

Interpretace identifikačních vět v **Mendoza2004** je velice podobná tomu, jak to dělá Weiss, ale je formulována v termínech aktuálních referenčních polí.

já:
Identifikace – ref (komp.1) = ref (komp.2)
Kromě toho existují případy tzv. «мнимой идентификации», když věta vypadá jako identifikační, ale ve skutečnosti spíše přisuzuje vlastnost. Jsou to věty typu *To byla na tom právě ta zvláštnost , to byl < ten vtip>* . (Dousková, Irena, Doktor Kott přemítá) Význam se nezmění, pokud tu větu přeformulujeme na *To bylo na tom právě zvláštní.* A v tom už nebude žádná identifikace. Nevím, co s tím, někam to zařadit.

X.2. Anotace – literatura z oblasti zpracování koreference v počítačové lingvistice.

V následující kapitole představíme přehled zahraničních prací a projektů anotace jmenné koreference a asociační anafory. U některých projektů provedeme srovnání anotačních principů a pravidel s naší

anotací. Toto srovnání používá typy a termíny naše anotace, které zavádíme až později v textu. Avšak takové srovnání pokládáme za dost přínosné, takže ať tam bude.

Chiarcos, Krasavina 2005

Manuál anotace koreference na RST Discourse Treebank (Carlson aj. 2003) a korpusu německých komentářů Postdam Commentary Corpus (Steed 2004). Anotace je rozdělená na dvě části – základní anotace koreference, teoreticky jazykově nezávislá, s omezeným počtem příznaků, lehce změňovatelná a adaptovatelná k novým cílům a zavedení/změně příznaků, vhodná pro budoucí automatické zpracovaní koreference a jiné experimenty z oblasti počítačové lingvistiky. Rozšířená anotace koreference je zpracovaná pro konkrétní jazyk (angličtinu a němčinu), má více ambiguity a méně přesnou sémantiku typů, je méně vhodná pro automatické zpracování, ale obsahuje z lingvistického hlediska více informace, může být tedy vhodnější pro lingvistické výzkumy.Anotace probíhá na textu, kde se nejdříve označí tzv. „markables“ - jednotky, které podílejí na anaforické návaznosti textu, hlavně NP nebo PP. Ty se dělí na primární a sekundární. V základní verzi anotace primární jsou osobní zájmena (vyjádřená v textu), určité a posesivní deskripce, pojmenované entity a názvy, pronominální advérbia; sekundární jsou hlavně neurčité deskripce, které se označí, pokud vystupují v textu jako antecedent v anaforickém vztahu. V rozšířené verzi k základním jednotkám přibývají tázácí, reflexivní a nulová zájmena, k sekundárním – propozice. Anotují se coreferenční anaforické vztahy a kataforické vztahy v rámci jedné věty. V rozšířené verzi se tyto vztahy klasifikují na modifikace (vyjádření jinými slovy, dodávání nové informace), synonymie, opakování stejné nebo skoro stejné NP (*der Kanzler ... der Kanzler ... Der Bundeskanzler*) a pronominalzaci. Rovněž v rozšířené verzi se anotují bridging vztahy, jejichž klasifikace se opírá na Gardent 2003. Při anotaci se postuluje řada konvenčních preferencí, jak se má anotovat některé typy případů, pokud existuje několik možností. Technický nástroj pro anotaci je MMAX. Anotují na textu (nikoliv na stromě) a na složkovém principu.

VIZ příloha X srovnání různých koncepcí

Vieira – Teufel 1997

(Renata Vieira and Simone Teufel, Penn Treebank korpus + WordNet, Edinburg, UK)
V (Vieira 1997) je představen pokus o automatické zpracování bridging vztahů na materiálu 20 článků z Wall Street Journal. Autoři projektu mají podstatně jiné rozdělení koreference na identickou a bridging, než používáme my pro anotaci PDT. Jako bridging se rozumí rovněž vztah synonymie a hyponymie - hyperonymie u koreferenčních jmen.
Autoři spravedlivě tvrdí, že pro automatické zpracovaní asociační anafory je třeba zapracovat dostatečný slovník. Pro tyto potřeby byla použita veřejně přístupná lexikální databáze WordNet (Miller, 1993). S použitím této databáze byly provedeny experimenty identifikace bridging anafory. Ukázalo se, že slovník pomáhá odhalit pouze 19 procent vztahů, označených ručně anotátory. Budou to především vztahy synonymické, hyponymické (hyperonymické) a vztahy typu část – celek. Neodhalují se především vztahy mezi pojmenovanými entitami a určitými deskripcemi (jako např. Mrs. Park -- the housewife and Pinkerton's Inc -- the company), anafora na větu nebo VP (Kadane Oil Co. is currently drilling two wells... -- The activity ...), složené deskripce, pro které je důležitý nejenom řídicí uzel, ale také všechny jeho určení (stock market crash -- the markets, and discount packages -- the discounts), sémantické vztahy návaznosti, důsledky, množiny-podmnožiny a některé jiné, které výrazně přispívají ke koherenci textu. Kromě toho nebyly odhaleny některé vztahy typu synonymickým, hyponymických a části - celku, které by teoreticky mohly být rozpoznány (Srov. např. a) Synonymy: new album -- the record, three bills -- the legislation; b) Hypernymy-Hyponymy: rice --the plant, the television show -- the program; c)Meronymy: plants -- the pollen, the house -- the chimney.).

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

MATE

MATE Dialogue Annotation Guidelines,

8 January 2000, Mengel, A., Dybkjaer, L., Garrido, J.M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C.

bridging

So-called bridging references (Clark, 1977) are expressions that denote objects only related to the denotation of their antecedent by (shared) generic knowledge. An example is the indicators in:

John has bought a new car. The indicators use the latest laser technology.

We are able to interpret the description the indicators because we know that indicators are a part of cars, and a car was mentioned in the first sentence. Some of the relations that may hold between a bridging reference and its antecedent include part-whole as in the example just seen, and element-set (as in The Italian team didn't play well yesterday until the centre-forward was replaced in the 30th minute). A bridging reference may also refer to the object filling a role in an event, whether implicitly or explicitly introduced, e.g. A young woman was attacked earlier this evening on Town Moor. The assailant was chased by a member of the public, but managed to escape. (A detailed survey of alternative classifications of bridging descriptions proposed in the literature can be found in Vieira (1998).)

Whether one is working on text or dialogue, the main problem in annotating anaphora is that almost every word in a text may be anaphoric (in the generalized sense discussed above) to some extent; hand-annotating all anaphoric expressions and all anaphoric relations is therefore impossible, except for small amounts of text. When designing a scheme for annotating anaphoric relations it is then necessary to identify the anaphoric expressions and relations more relevant for one's needs.

For instance, it is quite common to ignore first and second person pronouns when marking. It is not clear whether to mark appositions in noun phrases separately (as in "one of engines at Elmira, say engine E2 " or "The Admiral's Head, that famous Portsmouth hostelry "). Similarly, noun phrases in post-copular position can be problematic. For example, it can be argued that in (1.1) a policeman is clearly expressing a predicate, and therefore need not be marked, whereas in (1.2) (to be imagined being said while looking at the sky at night), both the planet on the left and Venus are clearly referring expressions; it's not so clear how to handle the president of the board in (1.3).

(1.1) John is a policeman.
(1.2) The planet on the left is Venus.
(1.3) John is the president of the board.

Massimo Poesio, 2004. "The MATE/GNOME Scheme for Anaphoric Annotation, Revisited", Proc. of SIGDIAL, Boston, April.

the scheme used for annotation of references to landmarks in the MapTask corpus - найти?

The Core Scheme In the most basic type of corefer ence scheme, only anaphoric relations between NPs are considered, and only identity relations. Schemes of this type can be implemented by having just one anaphoric relation, IDENT.

Про аннотацию части целого и подмножества вспоминает Renata Vieira and Simone Teufel результаты получаются очень неточными (у них там было мах 19%)

В MATE рассматриваются только ИГ, но не только прономинальные соединения

В проекте **GNOME**
After testing a few types of associative reference (Hawkins, 1978), we decided to annotate only three non-identity relations, as well as identity. These relations are a subset of those proposed in the 'extended relations' ver-sion of the MATE scheme: set membership (ELEMENT), subset (SUBSET), and 'generalized possession' (POSS), which includes both part-of relations and ownership rela-
tions.

То есть сначала они тоже взяли много отношений, а потом сузили их до меньшего количества - всего трех
К количеству отношений:
Restricting the range of associative relations The range of associative relations tested in GNOME is much narrower than those considered in DRAMA, but they can be annotated reliably, at least in the sense that very few disagreements are observed. Extending the range of relations to include, for example, attributes (e.g, I am not going to buy that. The price is too high. or situational associations (John entered a restaurant. The waiter approached him immediately) has proven difficult.

The identification of sentences, units and markables was done entirely by hand, without encountering particu-lar problems.

!!! Инструкции аннотаторам:
In order to achieve reliability on anaphoric annotation, the range of anaphoric phenomena considered was restricted in many ways. Apart from marking a limited number of associative relations, the annotators only marked relations between objects realized by noun phrases and not, for example, anaphoric references to actions, events or propositions implicitly introduced by clauses or sentences. We also gave strict instructions to our annotators concerning how much to mark. They were told to mark all identity relations, but to mark associative relations only if either
(i) no IDENT relation could be marked for the anaphoric expression, or
(ii) an IDENT relation with an entity not mentioned in the previous huniti. Furthermore, preferences were specified, e.g., for appositions: for example, *in Francois, the Dauphin*, the embedding NP would be chosen as an antecedent of subsequent anaphoric references, rather than the NP in appositive position.

!!!Результаты аннотации
We found a reasonable, although by no means perfect, agreement on identity relations. In a typical analysis (two annotators looking at the anaphoric relations between 200 NPs) we observed no real disagreements; 79.4% of these relations were marked up by both annotators; 12.8% by only one of them; and in 7.7% of the cases, one of the annotators marked up a closer antecedent than the other.

Limiting the relations did limit the disagreements among annotators on associative relations (only 4.8% of the re-lations are actually marked differently) but only **22% of bridging references** were marked in the same way by both annotators; 73.17% of relations are marked by only one or the other annotator. Reaching agreement on this infor-mation involved several discussions between annotators
and more than one pass over the corpus (Poesio, 2000a).

Причем они проводили аннотацию в два этапа - сначала вручную выбирали ИГ

Predicative NPs During the GNOME and VENEX annotations we realized that the recommendation not to mark predicative NPs makes it impossible to do markable identification automatically. In addition, it's often difficult to decide whether an NP is used predicatively or referentially, especially in languages like Italian where subjects in such clauses are often used predicatively (as in La soluzione e' questa).

Похоже на мои колебания **с нереферентными ИГ**:
One aspect of the markup scheme that needs revision is the placement of the semantic relation. One problem we observed in GNOME is that often the ambiguity is not simply between two possible antecedents each of which stands in the same relation to the anaphoric expression, but between two antecedents which stand in different relations. In the pharmaceutical texts, for example, it is of-
ten unclear whether a particular mention of the medicineunder consideration refers to the generic product, or to the particular instance that the user has in their hands. In this case, we would want annotators to mark the anaphoric expression as IDENT with one object, and ELEMENT of the other (ELEMENT is also used in GNOME for relations between instances and types), as follows, but this is not possible in either the original MATE scheme or in the GNOME markup scheme

Ambiguity Offering annotators the opportunity to annotate anaphoric ambiguity is essential, especially for annotations used to study linguistic phenomena, but raises serious theoretical and practical problems. A coreference chain containing such links becomes a coreference (directed) graph, in which each of the paths across the graph is a potential interpretation. While having multiple paths
is not a problem as far as evaluating the results of an anaphoric resolver (any path in the graph counts as a valid solution), it is a serious problems both for scripts attempting to ensure consistency (e.g., that all references to the same object are marked as either generic or non-generic– this is of course impossible when one of the possible antecedents is generic while the other isn't) as well for an-
notation tools (the problem is of course worsened when the tool only uses a single attribute to indicate membership in a coreference chain).

MUC

L. Hirschman. 1998. MUC-7 coreference task definition version 3.0. In N. Chinchor, editor, In Proc. of the 7th Message Understanding Conference.
MUC (Message Understanding Conference)
MUCCS was designed to encode information deemed useful for a subtask of information extraction, and the instructions provided to annotators were meant to ensure that all information provided by a text about a certain entity would be marked using a single device, the IDENT relation.

Т.е. все делается с целью дальнейшей автоматической эксцерпции информации из текстов. Однако для этой цели не всегда достаточно того, что у них есть (все фамилии президентов этой фирмы)

Based on experience in defining annotation schema for other phenomena, it is more important to preserve high inter-annotator agreement than to capture every possible phenomenon that could fall under the heading of "coreference". For this reason, the annotation scheme covers only the "IDENTITY" (or IDENT) relation for noun phrases; it does not include coreference among clauses, nor does it cover other kinds of coreference relations (set/subset, part/whole, etc.)

In keeping with having the coreference task support other information extraction tasks, we propose to place highest priority on preserving reasonable semantics for the equivalence classes. This means that two values (or instances) that are clearly distinct should NOT be allowed to merge into an equivalence class, even if this means not being able to mark all of the function/value or type/instance relations we might want to mark.

случай неверной цепочки с stock price
For example, in the sentence, "the stock price fell from $4.02 to $3.85", the stock price at one time is coreferential with $4.02, and at a later time with $3.85. However, if we make both $4.02 and $3.85 coreferential with stock price, we get "collapsing coreference chains" -- that is, we end up with *stock price*, *$4.02* and *$3.85* all in the same equivalence class -- which is counter-intuitive, and would prevent the IDENT relation from supporting, e.g., the Template Element task.
Thus, in the example above, we would mark *stock price* and the more recent value *$3.85* as coreferential, and leave *$4.02* in its own equivalence class, not marked coreferential with *stock price*. This means that the mark-up fails to capture some information (that *$4.02* is also a value, at an earlier time, of *stock price*), but this seems like a reasonable price to pay for preserving the semantics of the coreference equivalence classes.

что аннотируется: tried for an extensive but not exhaustive coverage of coreference phenomena:
- for now, only nouns are linked -- relations involving verbs are ignored: NOUNS, NOUN PHRASES, and PRONOUNS (personal including the possessive and demonstrative).
- Dates ("January 23"), currency expressions ("$1.2 billion"), and percentages ("17%") are considered noun phrases.
- A noun phrase is markable whether it is the object of an assertion, a negation, or a question.
- предикативы, ИГ в предикативном отношении: *Henry Higgins*, who was formerly *sales director for Sudsy Soaps*, became *president of Dreamy Detergents*
we have a sequence consisting of the extensional description *Henry Higgins* (which is the grounding instance), together with two intensional descriptions, *sales director for Sudsy Soaps* and *president of Dreamy Detergents*. In addition, there are two other extensional descriptions, *Sudsy Soaps*, and *Dreamy Detergents*.
in the example **Bill Clinton* is *the President of the United States**. we would record a coreference link between "Bill Clinton" and "the President of the United States". Coreference should NOT be recorded if the text only asserts the possibility of identity between two markables. In *Phinneas Flounder may be the dumbest man who ever lived*.
- First, second, and third-person pronouns are all markable (*There is no business reason for *my* departure", *he* added*. "my" and "he" should be marked as coreferential.)

- координация: we mark coreference between "The sleepy boys and girls" and "their" as follows: <COREF ID="1" MIN="boys and girls">The sleepy boys and girls</COREF> enjoy <COREF ID="2" REF="1" TYPE="IDENT">their</COREF> breakfast. In addition, the individual conjuncts are markable if they are separately coreferential with other phrases:
- сложная система, какие ИГ выбтрать для кореференции. Нам этого решать не нужно, у нас узлы и зависимости
- аппозиции: This identity of reference is to be represented by a coreference link between the appositional phrase, "the well-known emperor" and the ENTIRE noun phrase, "Julius Caesar, the/a well-known emperor": <COREF ID="1" MIN="Julius Caesar">Julius Caesar, <COREF ID="2" REF="1" MIN="emperor" TYPE="IDENT"> the/a well-known emperor,</COREF></COREF>. Appositional phrases are markable even when indefinite, e.g., *Ms. Ima Head, a 10-year MUC veteran*, *San Diego, one of America's finest cities*, An appositional phrase is also marked in the specifier relation, e.g., <COREF ID="1" MIN="job">The job of <COREF ID="2" REF="1">manager</COREF></ COREF>. However, appositional phrases are NOT marked when they are negative: *The criminals, often legal immigrants*, ...Appositional phrases are marked only when they constitute a separate noun phrase following the head. In written text, appositives are generally set off by commas

что не аннотируется
- Interrogative "wh-" noun phrases are NOT markables, e.g. "Which engine" and "Who"
- The relation is marked only between pairs of elements both of which are markables. This means that some markables that look anaphoric will not be coded, including pronouns, demonstratives, and definite NPs whose antecedent is a clause rather than a markable.
- Names and other Named Entities are all markables. A substring of a Named Entity, however, is not a markable. (Equitable of Iowa Cos. ... located in Iowa.)
-Date expressions recognized by the Named Entity task are also treated as atomic; components of a date are not separate markables. (*In a report issued January 5, 1995, the program manager said that there would be no new funds this year*. - нет отношения между *1995* и *this year*)
- к кореф. прилагательных - только если отсылают к НЕ или корню нормальной ИГ: Prenominal modifiers (e.g., *ocean drilling* in "the ocean drilling company") are markable only if either the prenominal modifier is coreferential with a named entity or to the syntactic head of a maximal noun phrase. That is, there must be one element in the coreference chain that is a head or a name, not a modifier. Thus the following instance is markable, because the prenominal modifier "aluminum" is coreferential with the head noun "aluminum" in the phrase "market for aluminum". *The price of *aluminum* siding has steadily increased, as the market for *aluminum* reacts to the strike in Chile.*
- нулевые элементы, перспроны: Assume that English has no zero pronouns; in other words, the empty string is not markable. In *Bill called John and spoke with him for an hour.* there is no relation between the implicit subject of "spoke" and "Bill".
- "который" *the movie which I saw* the relative pronoun "which" bears no markable relation to either "the movie" (the head to which the relative pronoun attaches) or to the implicit object of "saw" (the gap that the pronoun fills).

types or kinds, not to sets - между ними постулируется кореференция. У меня это выражено типом НР - если нереф., то значит тип. Примерно

3. Anotace rozšířené koreference na PDT

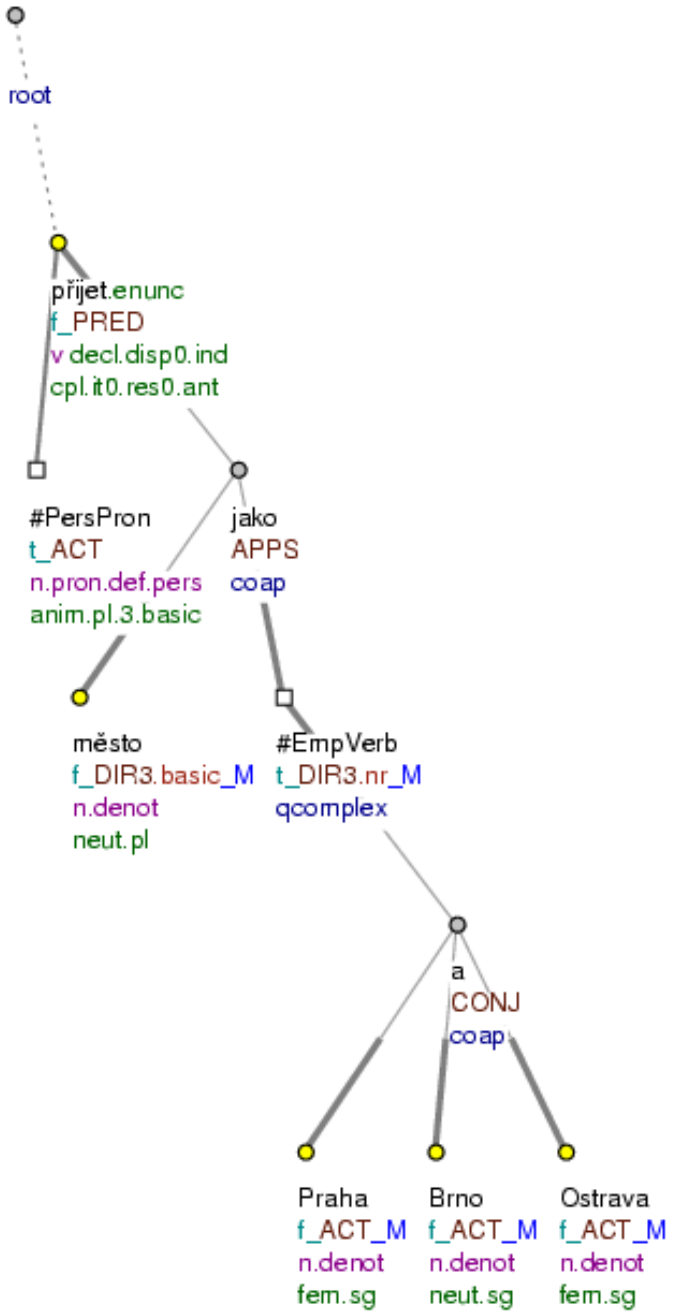
Rozlišujeme tři typy anotovaných vztahů: gramatická koreference (červená šipka), textová koreference (tmavomodrá šipka) a asociační/bridging anafora (bleděmodrá šipka). Gramatická koreference se týče koreferenčních vztahů uvnitř jedné věty, které jsou přesně determinovány gramatickými pravidly daného jazyka a jsou již kompletně oanotovány v původní verzi PDT 2.0. Textová koreference je anotována v původní verzi částečně (_VIZ_), bridging vztahy zavádíme nově. Syntaktická struktura anotovaných stromu PDT umožňuje zjištění některých koreferenčních vztahů automatický. Platí tedy princip, že pokud je koreferenční vztah zřejmý ze struktury stromu, tak ho neanotujeme.

Z těchto důvodů neanotujeme:

a) Vztah mezi jednotlivými členy **apozice** (Z podstaty apozice vyplývá, že jednotlivá pojmenování jsou koreferenční. Apoziční vztah je zachycen v tektogramatické stromě funktorem APPS a syntaktickou strukturou). Srov.

identická koreference: *Božena Němcová, autorka Babičky, slečna Sollárová, jinak.APPS slovenská malířka, ODS (Občanská demokratická strana)*

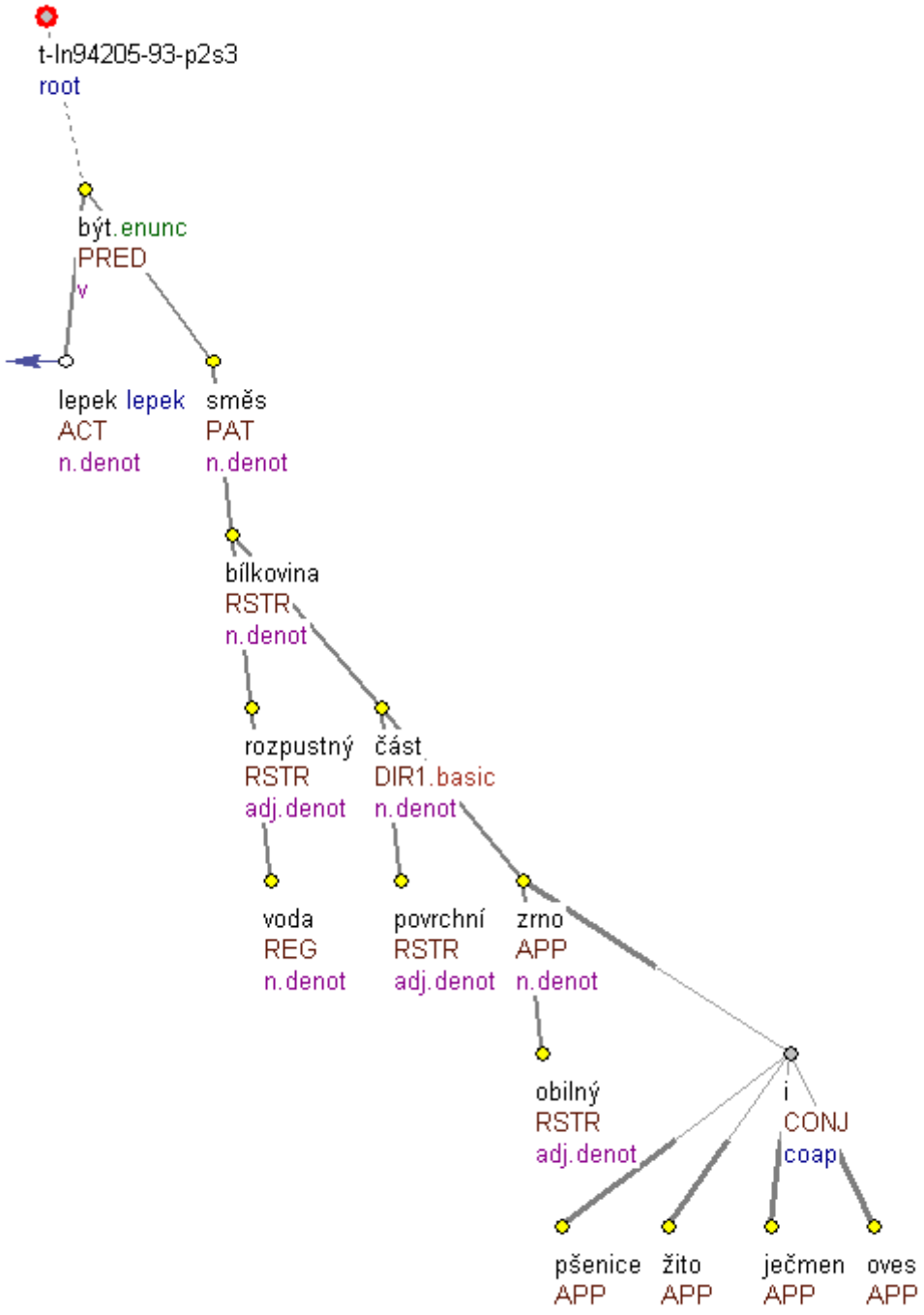
bridging: *Tomu odpovídala cílová místa - Kypr, Kréta, Malta. Přijeli do měst, jako Praha, Brno a Ostrava.*



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. Tento vztah je reprezentován v syntaktické struktuře stromu. Případnou koreferenční šipku vedeme na subjekt (resp. od subjektu).

Např. neanotujeme koreferenční vztah mezi uzly *lepek* a *směs* v následující větě.

> Lepek {coref_text na lepek v předchozím kontextu} je směs {žádná koreferenční šipka} ve vodě rozpustných bílkovin z povrchní části obilných zrn pšenice , žita , ječmene i ovsa .



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

(18/ln95047_061) My tady máme dost problematických dětí a ty kdyby se spolčily s Němci , to by nedělalo dobrotu .
(19/ln95047_061) Slyšeli jsme , že to {coref_text, typ_0 na „Němci“} jsou děti {žádná koreferenční šipka}, které místo výkonu trestu mají být tady v Košťanech . . . " říká obyvatelka obce a signatářka petice .

Jiná situace nastává v identifikačních větách, kdy oba členy predikačního vztahu mají vlastní referenci, identifikační syntaktickou konstrukcí se pak postuluje pouze identita těchto referentů. V takovém případě anotaci podléhají obě části predikačního vztahu, ale přesto neanotujeme koreferenci mezi subjektem a jmennou částí přísudku – jejich koreference je dána syntaktickou strukturou stromu a může být podle potřeby dodělána automaticky (_VIZ_ NĚKDY NA KONCI TECHNICKÉ POZNÁMKY, CO MÁ A JAK BÝT DODĚLÁNO AUTOMATICKY). Srov. např.

> Prvotní apoštolská církev byla chudá. Přesto i ona měla jakousi finanční organizaci, dokonce svého pokladníka. Problémem je, že tímto prokazatelně prvním křesťanským ekonomem {coref_text na "pokladník"} byl Jidáš Iškariotský.{žádná koreferenční šipka} Neblahé stigma Ježíšova zrádce {coref_text na "Jidáš"} jako by se nad církevním majetkem vznášelo dodnes.

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

V případě bridging vztahu tato konvence funguje trochu jinak a zatím se mi zdá, že ne úplně zřetelně. Momentálně aplikované pravidlo je, že MŮŽEME propojit subjekt s elementy části přísudku, pokud v přísudku je např. souřadná konstrukce, jejíž členy jsou vzhledem k subjektu v jednom z označovaných bridging vztahů, přičemž považujeme ten vztah za důležitě přispívající ke kohezi textu. V tom případě odkazujeme na uzel, který s antecedentem sémanticky souvisí. _VIZ_ k tomu poznámku v C.

3.A. Gramatická koreference

Gramatická koreference je takový typ koreference, kdy je možné určit antecedent na základě gramatických pravidel daného jazyka. Za gramatickou koreferenci se považuje koreference zvratných zájmen (*Sobě nedopřeje matka nikdy nic*), koreference vztažných prostředků (který, jenž apod.), koreference v recipročních konstrukcích (*Sultáni se vystřídali* {#Rcp.PAT} *na trůnu*), koreference u doplnění s dvojí závislostí vyjádřených slovesnou formou (doplňky apod. - *Mužstvo zůstává neporaženo*.PAT {#Cor.PAT} *i po tomto napínavém zápase.*), kontrola a kvazikontrola. Podrobněji k gramatické koreferenci _VIZ_ (Mikulova a kol. 2005, s. 935). Anotace gramatické koreference byla provedena částečně automaticky na celém korpusu PDT 2.0. Výsledky automatických procedur a další informaci jsou dokumentovány v (Kučová a kl. 2003). V anotaci rozšířené koreference zůstávají všechny původní šipky – na anotaci gramatické koreference se nic nemění.

A.1. Dodržování koreferenčního řetězce mezi gramatickou a textovou koreferencí

V původní anotaci koreference na PDT se snažilo o udržování nepřetržitého řetězce mezi šipkami gramatické a textové koreference. To znamená, že pokud A, B a C jsou po sobě následující jmenné

fráze, přičemž B je propojeno s A gramatickou koreferencí a C je textově koreferenční s A, šipka textové koreference vede od C k B. Srov.:

PRIKLAD Z TRAIN-2 NA PROPOJENY RETEZEC

Avšak tento princip nebyl dodržen pravidelně. Např. v souboru train-2 ze 302 podobných případů, koreferenční řetězec udržují pouze 242 případů (80%). Při anotaci rozšířené koreference jsme tuto nepravidelnost v anotaci původní pronominální koreference opravili automatickým mechanismem (_VIZ_ NĚJAKÁ TECHNICKÁ KAPITOLKA, KTERÁ JEŠTĚ NENÍ NAPSÁNA).

Při anotaci rozšířené identické koreference koreferenční řetězec rovněž udržujeme. Srov. např. v následujícím příkladě vztahujeme 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.

> (13/In94204_107) *Linka bezpečí 855 44 33* , *kterou* {coref_gram na „linka“}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.

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

Pozor! V případě asociační anafory se koreferenční řetězec nedodržuje. (_VIZ_)

B. Textová koreference

Základní princip textové koreference je identita referentů antecedenta a koreferujícího člena. Vztah koreference je symetrický (pokud A je koreferenční s B, B je koreferenční s A) a tranzitivní (pokud A je koreferenční s B je koreferenční s C, pak A je koreferenční s C).

Koreference se mohou zúčastnit NP v asertivních, otázkových i negovaných větách.

Na dnešní etapě anotace rozlišujeme původní pronominální textovou koreferenci a rozšířenou (hlavně substantivní) koreferenci.

Pronominální textová koreference je ručně anotována na celém korpusu PDT. Při anotaci se vyznačovaly vztahy:

- u osobních a přivlastňovacích zájmen pro 3. osobu. Tyto zájmena mají v tektogramatickém stromě jednotnou podobu #PersPron. Zájmena 1. a 2. osoby se nevyznačovala.
- u ukazovacích zájmen *ten, ta, to* v substantivní funkci.
- při aktuální elipse, kdy je do tektogramatického stromu doplněn nový uzel se zástupným t-lematem #PersPron (textová koreference tu není vyznačena v případech, kdy doplněný uzel zastupuje zájmeno 1. a 2. osoby). Při doplňování závislých valenčních doplnění všeobecným aktentem v podobě t-lematu #Gen případná koreference u těchto uzlů nebyla zachycena.

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

Rozšířená textová koreference se anotuje v současné době. _AKTUÁLNÍ INFO_

Další výklad je věnován pouze rozšířené textové koreferenci (dále jen textová koreference).

Textová koreference na velkou textovou vzdálenost

Textovou koreferenci anotujeme na vzdálenost nepřesahující 20 vět. V ostatních případech to neděláme především z důvodů předpokládaného velkého počtu chyb vzdálenější anotace. Anotátor jen ztěžka si vzpomene na antecedent, který se vyskytl v textu více než před 20 větami. Technicky je náročné při anotaci znázorňovat více než 20 předchozích vět (textové okno zabere příliš moc místa, TrEd je přetěžován předanotací a „padá"), takže ve výsledku chyb bude pravděpodobně víc než správně označených souvislostí. Tedy

Označujeme koreferenci v následujícím příkladě:

(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.

[…17 vět…]

(66/In94210_95) Existující kanalizační sítě {coref_text, typ_SYN na kanalizace"} by totiž podzemní chodbu vtlačily tak hluboko do země , že by jí jistě nikdo nepoužíval.

Avšak ji neoznačujeme v případě:

(6/In95047_061) Situace začala být přirovnávána k porevolučním snahám některých západních firem " odložit " na naše území za malý peníz toxické odpady .

[…44 věty…]

(50/In95047_061) Její zástupce ing . Šedivý však veškerou odpovědnost za krizovou situaci {žádná koreferenční šipka} odmítá .

Neanotujeme koreferenci ani v následujícím případě. Mezi větami je sice méně než 20 vět, ale nedá se tu mluvit o koreferenci. 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 {žádná koreferenční šipka} nahrazovat , ale doplňovat.

NBredace

B.1. Původní zájmenná koreference – některé případy 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 šipka 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 statusem 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 rozšířené koreference se pokusíme u párů s nespecifickou referencí a již označenou textovou koreferencí typ 0 měnit na NR (popis zkrátek a vysvětlení typu viz B.2.2.). 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**:
   substantivum[11] – 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*)[12]

ii. koreferující člen je vyjádřen **adverbiem**:
   antecedent (A)[13] – příslovce (*Praha – (#PersPron) – (ta) – (odtud) – tam*)

iii. jeden z členů páru je **adjektivum**. Tuto skupinu beru 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

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

[12] anaforický pár sloveso-substantivum je příbuzný z odkazem na celou situaci. Srov.

(4/In94207_84) Který čuně zase [#EmpVerb] ? !

(5/In94207_84) Řev {coref_text, typ_0 na #EmpVerb} rázu spíše symbolického

Avšak není to vždy. Srov.

(13/In94207_84) Lidi nežvýkají , to jenom krávy .

(20/In94207_84) Pravda o tom , že žvýkání {coref_text, typ_NR na „žvýkat“} pro žvýkání bylo odjakživa činností veskrze lidskou - kam paměť lidského rodu sahá .

[13] Antecedent ∈ {substantivum, zájmeno, elidovaný člen, příslovce}

případech se však bez takové anotace těžko obejdeme. Můžeme vyčlenit 2 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 šipku, 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). Srov.

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

Pokud však adjektivum nemá sémantický rys přináležitosti, koreferenci neanotujeme. Tak v následujícím příkladě anotujeme koreferenci u „podnikatelovy“ nikoliv však u podnikatelský. Srov.

(19/In94208_11) Tímto faktorem je podnikatel - inovátor , který se snaží o zisk , a proto logicky nemůže existovat ve stavu statiky , která nezná ani zisk , ani ztrátu .
(22/In94208_11) Podnikatelova {coref_text, typ_0 na „košťanský“} odměna , zisk , má však svůj původ nikoliv ve fungování , ale v rozbití stacionárního systému .
(25/In94208_11) Tento druh podnikatelské {žádná šipka} odměny je vlastně monopolní rentou a je dočasné povahy .

b) adjektivum je vytvořeno od pojmenované entity. Srov.

(96/In95047_061) Radní jednomyslně vyjádřili nesouhlas s přítomností chovanců společnosti Struktura v Košťanech .
(98/In95047_061) Ředitelka košťanské {coref_text, typ_0 na „Košťany“} základní školy Jarmila Hejduková byla jednou z iniciátorek podpisové akce .
(99/In95047_061) Pikantní detail v celé záležitosti je , že třinácti až čtrnáctiletí chlapci si dům v Košťanech {coref_text, typ_0 na „košťanský“} teprve upravovali .

Srov. také

(34/In94207_84) Významnou roli v dějinách žvýkačky sehrál mexický diktátor Antonio Lopez de Santa Anna .
(35/In94207_84) Poté , co byl v roce 1845 jako prezident svržen a na deset let vypovězen na Kubu , vydal se do New Yorku s jedinou myšlenkou - získat zpět vládu nad Mexikem {coref_text, typ_0 na „mexický“} .

V ostatních případech koreferenci na adjektivum neanotujeme. Srov.:

(24/In95047_061) Když se opat oseckého kláštera dověděl , jaké problémy s nimi jsou , další pobyt zakázal a němečtí hoši se museli i s vychovatelem z kláštera vystěhovat .
(25/In95047_061) Někteří se vrátili do Německa {žádná koreferenční šipka}, další přešli na faru v nedalekém Jeníkově a spojili se se skupinou , která tady byla již ubytována .

iv. členy koreferujícího páru jsou **číslovky**. Označujeme pokud vystupují v substantivní funkci a podílejí na kohezi textu. Srov.

(16/In94207_76) Připomenu , že po vstupu vojsk SSSR 21 . srpna 1968 na naše území jsem se zdržel (vraceje se z jugoslávských prázdnin) ve Vídni .
(21/In94207_76) Vzpomínám na takzvané zelené hranice zcela bezbariérové a na dosud nevídanou blahovůli zahraničních a našich celních a policejních orgánů už na jaře 1968 {coref_text, typ_0 na „1968“}…

v. **Sloveso** v roli koreferujícího člena vystupuje jen zřídka a v anotaci to zatím důsledně nezaznamenáváme. V některých případech je to ale vítáno. Srov.:

(68/In95047_061) Na převýchovu se pokud vím , posílali ti , kteří měli podle těchto zrůdných režimů nevhodný původ .
(70/In95047_061) Naše sdružení nepřevychovává {coref_text, typ_0 na „převýchova“}, ale snaží se vychovávat .

Srov. také část řetězce generických použití (*žvýkačky – nežvýkají – žvýkání – žvýkali – žvýkání*):

(2/In94207_84) Vybrané kapitoly z dějin žvýkačky
(13/In94207_84) Lidi nežvýkají {coref_text, typ_NR na „žvýkačka“}, to jenom krávy .
(20/In94207_84) Pravda o tom , že žvýkání {coref_text, typ_NR na „žvýkat“} pro žvýkání bylo odjakživa činností veskrze lidskou - kam paměť lidského rodu sahá .
(24/In94207_84) Se stejnou radostí však zamlčí , že Řekové často a s oblibou žvýkali {coref_text, typ_NR na „žvýkání“} kousky ztuhlé mízy mastikového keře, který se pěstuje především na ostrově Chios .
(25/In94207_84) Známý antický lékař a botanik Dioscorides psal v prvním století našeho letopočtu obsáhle o léčebném a hygienickém účinku žvýkání {coref_text, typ_NR na „žvýkat“}.

Ve funkci antecedenta sloveso vystupuje zcela běžně. (Srov. o tom dále v textu)

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 NP se specifickou referencí, kde koreferující člen není ani synonymum ani hyperonymum antecedenta (viz B.2.2.1.)

- **SYN** (od *synonymum*) – vztah mezi synonymními NP se specifickou referencí (viz B.2.2.2.)
- **ER** (od *hyperonymum*) – vztah mezi NP se specifickou referencí, kde druhý člen je lexikální hyperonym ve vztahu k prvnímu (viz B.2.2.3.)
- **NR** (od *nespecifická reference*) – vztah mezi NP s nespecifickou nebo generickou referencí (viz B.2.2.4.)

Poznámka: S typy SYN a ER (zvlášť s ER) se nepotkáme příliš často, pokládám však za nutné je mít, mimo jiné protože v některých pracích k automatickému zpracování koreference se s nimi pracuje jako s bridging. (např. Vieira 1997)

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 anotací je předvolený. Oba jeho členy mají specifickou referenci (odkazují ke konkrétnímu existujícímu, reálnímu referentu a objektu skutečnosti.) Srov např.

> (6/In9413_006) Jeho dojetí znásobila při vyhlašování přítomnost [...] pořadatelů soutěže - Českého manažerského centra v Čelákovicích .
> (7/In9413_006) Na letošním ročníku soutěže {coref_text, typ_0 na „soutěž“} se spolupodílí i Profit .
>
> (29/In9413_006) Začal jsem provozováním hospody, která {coref_gram, na „hospoda“}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 statusem (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, budeme to anotovat jako NP se specifickou referencí.

Jako coref_text, typ_0 označujeme také případy, kde opakovaní antecedentní NP je pouze částečné. Např. řetězce *společnost - akciová společnost - společnost Incheba*; *Vlček - ředitel J. Vlček - Jiří Vlček; ministr financí - ministr - tento ministr* atd.

Vyjádření identity pomocí textových identifikátorů. Příklad, kdy se se stejnou NP v anaforické pozicí 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 {coref_text, typ_0 na „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/In94204_107) Sedmiletý Péťa se půl roku neuvěřitelně trápil, že má AIDS. [...]
(29/In94204_107) #PersPron {coref_text, typ_0 na „Péťa“} Stále na to myslel, ve škole se už nedokázal soustředit. [...]
(34/In94204_107) Péťa {coref_text, typ_0 na #PersPron} skončil u Jany Drtilové.

Jako antecedent může vystupovat VP, celá věta nebo dokonce několik vět. V případě odkazu k několika větám použijeme tmavě červenou šipku speciální koreference (coref_special) typu segm. V ostatních případech šipka vede na řídící sloveso. Takovou koreferenci chápeme jako coref_text.

Pozor! Koreferenci neanotujeme u tzv. měřítek, např. uzlů jako #Percnt, bod apod. Srov. např.

Americký index obchodní důvěry odbytu a zaměstnanosti v příštích šesti měsících , se v srpnu snížil na 49.9 bodu , z 56.4 bodu {žádná koreferenční šipka}v červnu .
V dubnu byla jeho hodnota rovněž 49.9 bodu {žádná koreferenční šipka}.

Kataforický odkaz dopředu

Srov. také kataforický odkaz dopředu:

(11/cmpr9410_028) Tu nejvhodnější dobu {coref_text, typ_0 na „rok“}pan Hrabák propásl .
(12/cmpr9410_028) V osmdesátých letech se daly pořídit krásné věci za , viděno dneškem , ještě krásnější ceny.

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 replikách přímé řeči. Tento úkol se dá realizovat potom dodatečně, automaticky nebo částečně automaticky.

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 jiný typ koheze textu, který již přesahuje tekrogramatickou rovinu a patří spíše do roviny diskurzu. Při rozšířené anotaci tektogramatické roviny tento vztah nezachycujeme.

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.

Srov. také

(33/In94208_11) Dovozoval , že vývoj kapitalismu se historicky vyznačuje dvěma fázemi : Fází soutěžního kapitalismu a fází kapitalismu trustů .
(43/In94208_11) Schumpeter se ve svém posledním díle ptá : Který systém , kapitalismus {coref_text, typ_0 na „kapitalismus“}, či (44/In94208_11) socialismus , bude určovat budoucnost lidstva ?
K údivu , úžasu či ohromení většiny svých kolegů odpovídá jednoznačně : Bude to socialismus {coref_text, typ_0 na „socialismus“}.

B.2.2.2. Koreference synonymních NP (SYN).

Koreferující člen a antecedent jsou synonyma. 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.

Vztah SYN označujeme i přes #PersPron. Srov.

(57/In94207_84) Když o deset let později obrátil ke gumě pozornost louisvilleský lékárník John Colgan , existovala již

řada žvýkačkových milionářů (mezi nimi {coref_text, typ_0 na „milionář“} Adams) .
(60/In94207_84) Výsledek , který dostal obchodní jméno Taffy Tolu [], se ujal okamžitě , Colgan zavřel lékárnu a během krátké doby se přidal k milionářskému klubu {coref_text, typ_SYN na #PersPron } .

Také vztah SYN může vést na kořen souřadné nebo seznamové struktury. Srov.

(96/In94207_84) Jde především o zdravotní (či pseudozdravotní) námitky proti žvýkání .
(97/In94207_84) Tento boj {coref_text, typ_SYN na „či“, sémanticky však samozřejmě na námitky } začal již počátkem století značkou Dentyne .

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

Z koupelny tiše bzučela automatická pračka a z rádia 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 city k té rajdě si nechal přerůst přes hlavu tak , že přišel o plody celé poloviny svého poctivě pracovitého 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 vtipkováním 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)

Za synonymickou koreferenci označujeme rovněž vztahy s pojmenovanými entity, i když to není úplně samozřejmé (viz jinak u Vieira 1997) Srov.

Nbpříklad

Textovou koreferencí typu SYN spojujeme rovněž uzly, z nichž jeden je vyjádřen zkratkou (např. *ČR - Česká republika, ODS - Občanská demokratická strana* apod.) Srov.

O odpočtu DPH
Podle novely zákona o dani z přidané hodnoty { coref_text, typ_SYN na DPH } se letos stanu plátcem daně .

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. následující příklady.

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

> (128/In94207_76) Původně měly vznikat jako rozmluva na diktafon , tento způsob práce {coref_text, typ_ER na „rozmluva“}však L . Fuks nepřijal a začal přinášet již hotové texty .

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)

Koreferenci typu ER označujeme v párech sloveso (situace) – obecný název té situace. Srov.

> (11/cmpr9410_001.t)Jistotu v tomto směru dávají nejnovější kroky vlády SR , která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční [provinencie]provenience .
> (12/cmpr9410_001.t)Byť má na tento krok {coref_text, typ_ER na „rozmluva“} určité právo (jako člen GATT) , v daném okamžiku však vyznívá jako tvrdé politické rozhodnutí vlády , která se snaží velice rezolutními administrativními kroky zredukovat mnohamilionové pasívum v obchodní výměně s ČR .

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

Nespecifickou referenci zaznamenáváme u:

- uzlů. závislých na „konteinerech“ (sklenice mléka)
-

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 pravděpodobné jiné chování synonymních a hyperonymních koreferujícího členu – častější užívání ukazovacích zájmen, možné jiné pozice ve větě (vždy výrazně tématické) apod.

V textu koreferenční řetězce typu 0 a typu NR se mohou prolínat, tj. některé hrany jednoho řetězce mohou být označeny NR, jiné však 0. Srov. příklad dlouhého hypertématického řetězce (je tu trochu zmatek, částečně proto to tam uvádím):

> (24/In95047_061) Když se opat oseckého kláštera dověděl , jaké problémy s nimi {specifická reference, coref_text, typ_0 na předchozí NP}jsou , další pobyt zakázal a němečtí hoši {specifická reference, coref_text, typ_0 na #PersPron}se museli i s vychovatelem z kláštera vystěhovat .
> (26/In95047_061) Podobné zkušenosti jako v Oseku potvrdil i farář v Jeníkově pan Matfiak .
> (27/In95047_061) I tady si prý chlapci {specifická reference, coref_text, typ_0 na NP v předchozím kontextu}, kteří měli být vychováváni na faře , užívali děvčat a svobody .
> (28/In95047_061) Posledním místem , kam byli chlapci {specifická reference, coref_text, typ_0 na NP v předchozím kontextu} firmou Struktura umístěni , byl bývalý dům dětí a mládeže v Duchcově .
> (29/In95047_061) Také lidé z okolních domů si stěžovali na hluk , výtržnosti , aroganci a proudy holek , které se za kluky {specifická reference, coref_text, typ_SYN na „chlapec“} táhly .
> (30/In95047_061) Podle názoru některých z nich jde o německou polepšovnu .
> (31/In95047_061) Bylo jim však divné , že chlapce {specifická reference, coref_text, typ_SYN na „kluk“} nikdo nevede , nehlídá ani nevychovává .
> (32/In95047_061) Kdo to {specifická reference, coref_text, typ_0 na „chlapec“} vlastně je ?
> (33/In95047_061) Německé chlapce {nespecifická reference, coref_text, typ_NR na „ten“}jsme již nezastihli .

(34/In95047_061) Duchcov byl posledním místem , odkud #PersPron {specifická reference, coref_text, typ_0 na „chlapec“}byli těsně před naším příjezdem odvezeni zpět do Německa .

Srov. také

(5/In94206_38.t) Dodal, že loni podalo tuto žádost 200 odsouzených {specifická reference} .
(6/In94206_38.t)Praktické předávání však začalo až letos v červnu , kdy bylo předáno 16 odsouzených .
(7/In94206_38.t)Další dva budou převezeni počátkem září .
(8/In94206_38.t)Malý počet předaných osob {coref_text, typ_NR na „odsouzený“ v 5} je podle něj způsoben především administrativními problémy .

??? 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 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 šipky 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).

Vybíráme tady první variantu.

Srov. také:

(29/cmpr9410_028) Unikát i za statisíce si kupce vždy najde , byť to trvá zpravidla déle .
(30/cmpr9410_028) Dříve by asi nezaváhala Národní galerie {žádná koreferenční šipka} , ale dnes její rozpočet často nestačí .

NP „Národní galerie“ má specifickou referenci, NP „kupce“ – nespecifickou. Je tam přitom dost jasná identita. Necháme to nepropojeně.

Další bod této problematiky je řešení mezi generickou a specifickou NP, které není vždy jednoznačné. Srov. NP „prací prášek“ v následujícím příkladu. Generická reference???

> U detergentu Toto jsme například řešili problém s udržením stálé kvality , protože jednotlivé partie byly nevyvážené . Investovali jsme dva miliony korun do nákupu pásových vah , zpřesnili dávkování a jakost pracího prášku stabilizovali .

B.2.3. Problématické případy označování textové koreference

B.2.3.1. 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 {žádná šipka} zdarma .
> (49/In94204_107) Podle zkušeností ze zahraničí se dá předpokládat , že největší frekvence telefonátů { coref_text, typ_NR na „volat“ } 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 {žádná šipka} 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?

Srov. také

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

(32/In94210_95) Paláce {žádná šipka} neznamenají přepych
(36/In94210_95) Ač se to na první pohled nezdá , obývání klasických renesančních a barokních paláců { coref_text, typ_NR na „palác“ } s velikými , řetězovitě propojenými místnostmi není žádné terno .

V páru dvou paláců (22) a (32) koreferenci neoznačujeme, protože v (22) NP „palác“ má specifickou referenci a v (32) – generickou, tedy (22) a (32) nejsou koreferenční. Vztah mezi NP v (32) a (36) označíme jako textovou koreferenci s typem NR.

V následujícím případě vyčleňujeme dva (nikoliv jeden!) nereferenční řetězce propojené textovou koreferencí: *míza – guma* (38 a 48) a 4x *chicle* s generickou (38, 41, 42) a specifickou (48) referencí.

(38) Tak jako každý Mexičan , i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle (prý z mayského slova tsictle) , a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku .
(41) Santa Anna má chicle { coref_text, typ_NR na poslední „chicle“ v předchozí větě} a Adams technické schopnosti .
(42) Asi rok se Adams a jeho nejstarší syn snažili - chicle { coref_text, typ_NR na „chicle“ v předchozí větě }vařili , čistili , přidávali množství různých látek a míchali s pravým kaučukem .
(44/In94207_84) Když asi po roce své úsilí vzdali , rozhodl se Adams , že vše , co mu z chicle { coref_text, typ_0 na „chicle“ v předchozí větě }ještě zbylo , hodí do řeky .
(48/In94207_84) Vzpomněl si totiž , jak Santa Anna čas od času uloupl kus gumy { coref_text, typ_NR „míza“ v (38)}, strčil do pusy a žvýkal .

Koreferenci typu NR neoznačujeme pokud extence (dosah, možné denotáty) daných NP mají různý dosah, čili nejsou koreferenční, i když o nereferenčních jmenných frázích to není úplně logické tvrzení. Např. v násl. příkladě jsou dvě generické jména, ze kterých druhé je „specifičtější“ než první, tj. odkazuje na omezenější množinu, třídu denotátu. Srov.

(17) Stali jsme se také [dodavatel]dodavatelem Unileveru a dokázali splnit jeho zvýšené požadavky na kvalitu .
(25) U detergentu Toto jsme například řešili problém s udržením stálé kvality (tady – kvality pouze detergentu), protože jednotlivé partie byly nevyvážené .

Srov. také následující: poplatek nemá materiální denotát. Stačí to pro to, aby daná NP byla generická? Momentálně označeno jako 0

Milionový poplatek za vydání osvědčení , které umožňuje vést lékárnu , zakázalo vybírat Ministerstvo pro hospodářskou soutěž .
Tento poplatek odhlasovali její členové na svém druhém sjezdu v říjnu 1992 .

B.2.3.2. Koreference abstraktních jmen

Další problematický bod v rozlišování specifické a nespecifické reference jsou **abstraktní jména**.

Přehled literatury k tématu

Základní, jednoduchá definice: *Konkrétní*, jsou ta, která představují nějaké hmotné věci, na které si můžeme sáhnout, např. strom, kámen, papír, vlasy... Naopak *abstraktní* představují slova, na která si sáhnout nemůžeme, např. pocit, strach, láska, představivost... (Internet)

Rozlišení abstraktní a konkrétní NP není jednoznačné, tato opozice je spíše graduální (Чернейко 1997). Velmi obecně jako abstraktní rozumíme NP myšlenkové, pojmové, zatímco konkrétní mají předmětný věcný obsah. Rozdělení na abstraktní a konkrétní lexiky je základní.(srov. už Frege 1892). Obě třídy (abstraktní a konkrétní) jsou však dost dynamické a není vyloučeno, že u některých jmen nebudeme přesně vědět, kam je zařadíme.

Přesná kritérie rozdělení jmen na abstraktní a konkrétní nejsou. Můžeme vyčlenit několik principů, podle kterých to rozdělení většinou provádíme.

1. referenční princip: konkrétní odkazují na reální materiálně hmatatelnou věc, abstraktní takový denotát nemají.
2. formálně sémantický princip: abstraktní jsou taková jména, která popisuj vlastnosti, stavy a vztahy věcí zvlášť od jejich hmotných nositelů.
3. sématnický: abstraktní jsou jména, která mají širší význam ve srovnání s jinými slovy, a která jsou propojená s těmito slovy vztahem třída – představitel
4. syntaktický: abstraktní slova jsou častější predikátová

Ju.S. Stepanov dělí jména na denotátní a signifikátní. Denotátní slovní zásoba se směruje k označení reálních předmětů vnějšího světa, denotátů, zatímco signifikátní slovní zásoba spíše pojmenovává pojmy, signifikáty. (Stepanov 2004:59) K denotáním patří také obecné názvy, které se determinují výčtem součástí podle principu „část – celek“. Obecný termín je názvem určité situace, závislé termíny vytváří tématickou třídu, jako např. *Oblečení* (obecný termín) – *sukně, košile, ponožky* apod. (závislé termíny). Signifikátní jména jsou např. *zvíře* jako obecný název pro množinu *vlk, kráva, kůň* apod., mají strukturní vztyhy třída – jednotka, elementy této třídy mohou vždy zaměnit svůj hyperonymum.

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

Н.Д. Арутюнова исходя из синтаксических возможностей имени (в частности степени его синтаксической свободы), разделяет субстантивы на три класса: имена лица/ не лица, имена конкретно-предметного и абстрактно-событийного значения. Н.Д. Арутюнова приходит к выводу, что «чем более предметно значение существительного, тем затруднительней для него непосредственное включение в систему форм предложения» (Арутюнова 1976).

E.V.Padučeva má několik článků věnovaných referenčním schopnostem odpredikačních (dějových) substantiv a jiných jmen s nepředmětovým významem. Její základní myšlenka je v tom, že na rozdíl od predikátů, které nemají vlastní referenci, propoziční komponenty referují, a to na situaci, kterou pojmenovávají. Typ reference propozičních komponentů se definuje na základě několika parametrů: odkaz na situaci nebo na fakt(možnost). Další bod klasifikace je modalita (reální a neutrální)
V Padučeva (1986) provádí klasifikace propozicí podle jejich schopnosti referovat na fakt nebo na situaci.

Literatura k řešení:

Lezin2007 při realizaci projektu automatického vyhledávání referenciálního propojení textu zahrnuje abstraktní NP spolu s generickými a predikativními NP do jedné skupiny „třídy objektů“ a řeší je zvlášť od jmenných frází se specifickou referenci. Navíc zvlášť vyčleňuje skupinu „třída objektů aktuální pro daný diskurz“, kam zapadají pořád ještě generické NP, které jsou však o něco více specifikované pro účely daného textu. Srov. např. „*Žena nesmí do velké politiky*“ (skupina „třída“) vs. „*Žena v naši společnosti nesmí do velké politiky*“ (skupina „třída aktuální pro daný diskurz“)

* * *

Řešení je většinou dost arbitrární. V anotaci se však budeme snažit rozlišovat specifickou a nespecifickou referenci i u abstrakt. Děláme to zatím tak, že pokud daná NP má jasně generickou referenci, dáme ji typ NR. Pokud tomu tak není, označíme ji jako coref_text, typ 0. 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 šipka}a děsivých představ.

I když se zdá divné propojovat tyto NP textovou koreferencí, je to pořád stejný strach. V daném příkladě má antecedent ještě navíc funktor CPHR, což také „ochuzuje“ jeho referenční status. Avšak vzhledem k tomu, že strach je v obou kontextech stejný, budeme ho spojovat koreferencí.

Srov. také

> (33/In94208_11) Dovozoval , že vývoj kapitalismu se historicky vyznačuje dvěma fázemi : Fází soutěžního kapitalismu a fází kapitalismu trustů .
> (43/In94208_11) Schumpeter se ve svém posledním díle ptá : Který systém , kapitalismus {coref_text, typ_0 na „kapitalismus“}, či (socialismus , bude určovat budoucnost lidstva ?
> 44/In94208_11) K údivu , úžasu či ohromení většiny svých kolegů odpovídá jednoznačně : Bude to socialismus {coref_text, typ_0 na „socialismus“}.

Srov. ale s generickou referenci:

> (19/In94208_11)Tímto faktorem je podnikatel - inovátor , který se snaží o zisk , a proto logicky nemůže existovat ve stavu statiky , která nezná ani zisk {coref_text, typ_NR na „zisk“}, ani ztrátu .
> (27/In94208_11) Na konci tohoto difusního procesu se systém vrátí ke statické rovnováze , v níž nebudou opět ani zisky {coref_text, typ_NR na „zisk“}, ani ztráty .

B.2.3.3. Koreference dějových jmen

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

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

B.2.3.4. Problematické páry NP se specifickou referencí

S principiální otázkou označovat nebo neoznačovat koreferencí v daném konkrétním párů 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 budeme 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.

Z označování koreference mezi dvěma členy páru automaticky neplyne, že mají mezi sebou anaforický vztah, i když tam většinou je. Je to zřetelné na příkladech s jmennými frázi, které se vyskytují ve stejném textu ale v různých diskursivních jednotkách. Srov.:

(2/In95047_061)Před několika týdny zaplnily stránky regionálních deníků i celorepublikových časopisů články , jejichž titulky " Jugend prý nabízí dětem alkohol a svádí patnáctileté dívky " nebo " Většina obyvatel by zřejmě mezi sebe problémové děti , které k nám vozí na převýchovu německá církev , nepřijala " a " Němečtí problémoví odešli z Košťan " naznačovaly odhalení skandálu .
(11/ In95047_061) V obci Košťany na Teplicku (coref_text, typ=0 na „Košťany“)ještě chlapci ani nebyli , ale místní již dali dohromady petici : " My rodiče dětí základní školy Košťany protestujeme proti umístění ubytovny pro potrestané německé chlapce .

V podobných příkladech koreferenci nemůžeme neoznačit (referují jednoznačně na stejný mimojazykový objekt), avšak je jasné, že anafora to také není.

Srov. také (mezi replikami je 11 vět):

(4/In95047_061) Podle těchto zpráv nějaká firma na naše území umísťuje německou delikventní mládež , která zde páchá kriminální činy a ohrožuje starousedlíky .
(15/In95047_061) V Košťanech totiž zakoupila dům firma Struktura (coref_text, typ=0 na „firma“), která se u nás rozmísťováním německých chlapců zabývá .

Sporný příklad (koreference specifické a autonymní NP)

Je sporné, jestli máme označovat textovou koreferenci v následujícím případě:

> (20/cmpr9410_028) Kdo by si třeba v roce 1898 pomyslel , že za pouhých 600 franků neprodané Cézannovo zátiší {specifická reference}, které jakýsi znalec pohrdlivě označil slovy " křivé ovoce {autonymní použití, žádná koreferenční šipka} v kácejících se nádobách " , bude za dvě desetiletí ceněno na [300 000]300000 franků .

Nabízím to neoznačovat, částečně je ta informace obsažena v syntaktické struktuře věty.

B.2.3.5. Dvě místní určení vedle sebe (*tady v Praze, u nas doma* apod.)
Pokud v tektogramatickém stromě jsou dvě místní (možná i jiné, zatím jsem se nenarazila) určení jako sestry a nejsou přitom ve vztahu apozice, propojíme je textovou koreferencí postupně po sobě. Srov.

> (42/In94207_76) Na stůl přinášel kuchyni studenou , chlebíčky , uzeniny , šunku a západoněmecké sýry mnoha druhů a zde jsem žasl nad jejich kvalitou , kterou jsem z domova nepředpokládal .
> (46/In94207_76) Při tlumeném světle přicházela na přetřes politická situace u nás (coref_text, typ=0 na „domov“) doma (coref_text, typ=0 na „#PersPron“) .

Lingvistická zajímavost – anafora na neobligatorní a nevyjádřené místní určení:

> Jde o významný vztah nepřímé úměry - čím vyšší počet živností (*v regionu*), tím relativně nižší nezaměstnanost v daném regionu .

samozřejmě tady žádnou šipku nikam nevedeme. Prostě příklad na anaforu nikam.

B.2.4. Nejednoznačný výběr antecedentu

B.2.4.1. K otázce výběru antecedenta v případě apoziční skupiny:
Pokud koreferující uzel odkazuje na apoziční spojení, koreferenční šipka vede na spojku. Je to arbitrární a poněkud nepřehledne řešení – v řetězcích pak budeme mít náhodou spojku, o které bez kontextu neuvidíme, co vlastně spojuje. Avšak je to pravděpodobně jediný způsob zachovat jednotnost anotace. Srov.

> (35/In95047_061) Zastihli jsme tady pouze jediného vychovatele - pana Fuchse .
> (37/In95047_061) Podle pana Fuchse {coref_text, typ_0 na „#Dash“} nejde o žádné kriminálníky ani delikventní mládež .
>
> (72/In95047_061) Tato slova řekl Raimund Strathman , muž z Evangelické vesnice mládeže , která má sídlo v Rensburgu v německé spolkové zemi Šlesvicko - Holštýnsko .
> (82/In95047_061) Jsme domov mládeže , váš ' děcák ' , nikoliv ' pasťák ' , " říká R . Strathman {coref_text, typ_0 na #Comma, nikoliv na „Strathman“} .

B.2.4.1. K otázce výběru antecedenta v případě koordinační skupiny:
Pokud jde o identickou koreferenci, šipku vedeme na spojku nebo od ní (v případě bridging vztahů je to jinak – viz C.3.3.).

(34/cmpr9410_028) Vznik moderního umění se spojuje s rokem 1907 , kdy byla založena populární Osma , a v kubistickém a fóbistickém duchu malují Filla , Kubišta , Špála a další .
(35/cmpr9410_028) Ceny jejich {coref_text, typ_0 na spojku „a“} obrazů šplhají do statisíců a dobře se prodávají v cizině .

B.2.4.3. K otázce dvojího odkazování (identická koreference) - ???:
Setkala jsem se s případem, který se asi bude opakovat a s kterým si neumím moc poradit. NP s nespecifickou referencí má více než jednu možnost textového odkazování. Srov. např.:

(26/In94207_84) Pro historii žvýkací gumy , jak ji známe dnes , se však musíme přenést na jiný kontinent .
[…27-28: *výklad o tom, jak Mayové vyráběli žvýkačku…*]
(29/In94207_84) Mízu stromu sapodilla (achras sapota) sklízeli a upravovali systémem , který se používá dodnes .
(30/In94207_84) Kůru stromu nařízli do tvaru písmene v a do špičky řezu umístili nádobu , do níž šťáva ukapávala .
(31/In94207_84) Získanou mléčnou gumovitou látku pak čistili , vařili .
(32/In94207_84) Teprve výsledný substrát byl hoden žvýkání .
(33/In94207_84) Zrodila se žvýkačka{coref_text, typ=NR na „guma“, coref_text, typ=NR na „substrát“} .
(38/In94207_84) Tak jako každý Mexičan, i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle … .

Zdá se logické propojit jak „substrat“ a „žvýkačka“ v (32) a (33), tak i NP „žvýkací guma“ a „žvýkačka“ v (26) a (33), protože větou (33) se končí výklad o její výrobě. Vzniká tedy otázka, jestli máme tu označovat dvě tmavomodré šipky s identickou koreferencí. Je to trochu proti logice koreferenčních řetízku, ale odpovídá kohezi textu. V daném případě jsem zatím označila oba vztahy.

V druhé posloupnosti (*míza – šťáva – látka … míza*) je situace jiná. Máme koreferenční řetězec *míza → šťáva → látka* ve větách (29)-(31), potom ale následuje NP „míza“ v (38), která odkazuje k „míza“ v (29) a tedy podle pravidel textové koreference koreferenční se „*šťáva*“ a „*látka*“. Je ale jasné, že na ně neodkazuje. Avšak neanotujeme pouze anaforu, ale koreferenci (viz k tomu poznámka výše). Co tedy máme dělat, abychom zachovali důslednost anotace? V daném konkrétním případě bych „míza“ na „míza“, pak třeba vymyslím lepší pravidla. *U těch NR je to vždycky problém….*

B.2.4.4. Spojení se slovy s funkcí „kontejneru“
Ve spojení se slovy s funkcí „kontejneru“ (spousta, řada, milion apod.) šipka vede na kontejner. Srov. ne zcela intuitivní příklad:

(43/In94207_76) Ale přitom hostitel otevíral láhve alkoholu .
(44/In94207_76) Byla to {coref_text, typ_NR na „láhev“} vína a německé alkoholy tvrdé .

Křesťané se modlili za usmíření národů...
Více než tisícový zástup křesťanů z různých sborů a církví českých zemí a delegace křesťanů {coref_text, typ_NR na „křesťan“, funktor APP} z Německa se v sobotu na vrchu Radobýl u Litoměřic modlil za smíření mezi Čechy a sudetskými Němci .

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 typů, přičemž tento výběr je založen na literatuře z oblasti počítačové lingvistiky.

Preference! Pokud můžeme vybírat, odkázat na (třeba i vzdálenější v textu uzel) identickou koreferencí, nebo vztahem typu bridging, vždy volíme identickou koreferenci.

Přehled literatury k tématu bridging

Clark1977, Gardent2003
Gardent2003 uvadí následující 13 typů

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 dětmi 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 v soutěži podnikatelů {má specifickou referenci} .
(101/In9413_006) K různým soutěžím {bridging, typ=contrast, na „soutěž“} mám výhrady.

Ale:

(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! Oprotí identické koreferenci (gramatické a textové – viz A), pomocí bridging-vztahů propojujeme pokud možno jenom lexikálně vyjádřené autosémantické uzly (žádné „který“ a rekonstruovaný „PersPron“). V některých případech to nejde, potom můžeme odkazovat i na lexikálně nevyjádřený uzel. Srov. např.:

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

Dodržování koreferenčního řetězce mezi gramatickou-textovou koreferencí a asociační anaforou

V případě asociační anafory koreferenční řetězec neudržujeme a propojujeme jmenné fráze podle významu. Srov. např.:

(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 .

C.1. Typologie bridging vztahů

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

- **FUNCT** – vztah individual – function, viz C.1.
- **PART** – vztah část-celek (oba směry), viz C.2.
- **SET** – vztah množina-podmnožina (element množiny) (oba směry), 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.

Pozor! Bridging vztah neoznačujeme pokud uzly jsou spojené jednou závislostní šipkou, přičemž závislý uzel má funktor aktantu APP nebo AUTH. Takto spojené uzly mají mezi sebou velice často sémantické vztahy, z kterých mnoho se dá zařadit mezi vztahy typu bridging. Avšak tyto uzly už jsou dostatečně spojené závislostní šipkou a zatěžovat strom dalšími šipkami v tomto případě se nám zdá zbytečné. Srov. páry typu *obyvatelka obce* (vztah PART, funktor PAT), *opat kláštera* (vztah FUNCT, funktor APP), *starosta obce* (vztah FUNCT, funktor PAT), *člen výboru parlamentu*(vztah PART, funktor APP), dílo Wagnera (funktor AUTH), *jejich obrazy* (funktor AUTH) apod.

??? Nejsem si úplně jistá, jestli se toto pravidlo má dodržovat pravidelně. Občas bych bridging vztah ráda označila i když zapadá do této neoznačovatelné skupiny. Pokud původní nepronominální antecedent je v jiné větě??? Srov.

(81/In94207_76) Kdykoliv jsem navštívil Mnichov nebo jím projížděl , neopomněl jsem zajít do pověstné mnichovské pivnice Bierbräukeller .
(82/In94207_76) Její obrovská hala {bridging, typ=PART, na „pivnice“}, kde se sedělo u dubových stolů a bavorské pivo se nalévalo do litrových korbelů s příklopkami .

Bridging vztahy většinou odkazují dozadu anaforicky. Jsou ale výjimky. Srov. příklady bridging katafory:

Ministr Karel Dyba : Vzhledem k očekávané poptávce a dané sumě , která je k dispozici ze státního rozpočtu , byly v nových programech Ministerstva hospodářství pro rok 1994 provedeny následující změny {bridging, typ=SET_SUB, na kořene stromů následujících vět} :
Za a) . S výjimkou programu Region a Aeskulap byla snížena sazba příspěvku na úhradu úroků o 1 - 2 % proti roku 1993 .
Za b) . --> Byla zkrácena doba od podpisu úvěrové smlouvy k registraci žádostí o podporu z 1 roku na 6 měsíců .
Za c) .Byly zrušeny cenově zvýhodněné záruky za úvěr u jednotlivých programů a vyčleněny do nového programu Záruka .
Za d) . Byl zrušen příspěvek na úhradu úroků pro obchodní činnost s výjimkou programu Start .

V případě bridging katafory pořadí vztahů typu SET, PART a FUNCT označujeme podle lineárního pořadí.

C.1.1. bridging-vztah *individual – function* (FUNCT)

Vztah se dá nejlépe ukázat na příkladech. Teoretický komentář později.

Srov. *fara - farář*

(25/In95047_061) Někteří se vrátili do Německa , další přešli na faru v nedalekém Jeníkově a spojili se se skupinou , která tady byla již ubytována .
(26/In95047_061) Podobné zkušenosti jako v Oseku potvrdil i farář {bridging, typ=FUNCT, na „fara“} v Jeníkově pan Matfiak.

Vztah FUNCT funguje v obou směrech (funkce před objektem nebo po něm). Srov. pár *ministr – ministerstvo*:

(52/In95047_061) Mezitím starosta obce Košťany Jindřich Abrhám požádal o pomoc člena branně - bezpečnostního výboru parlamentu poslance Čapka (Levý blok) , který se obrátil na ministra Rumla s žádostí o prošetření činnosti firmy Struktura .
(53/In95047_061) Ministerstvo {bridging, typ=FUNCT, na „ministr“}vnitra však odpovědělo , že nejsou známy žádné negativní informace o její činnosti .

Bridging vztah typu FUNCT v páru ministr Karel Dyba - ministerstvo

řešíme jako ministr Karel Dyba - ministerstvo {bridging, typ=FUNCT_P, na „ministr“} Postupujeme tak ze dvou důvodů:

1) u bridging nemusíme dodržovat řetězec, takže není třeba vést násilně na řídicí uzel;
2) bridging vztahy se snažíme pokud možno označovat u slov, která ty významy mají v lexikální sémantice

Pokud však v paru Karel Dyba - ministerstvo není přímo uvedeno, ze Dyba je ministr, a víme to jenom ze znalosti světa, žádnou koreferenci neoznačujeme.

Kontextově podmíněný vztah FUNCT

Méně jednoznačný, avšak ještě zapadající do vztahu funkce je následující příklad:

> (24/In95047_061) Když se opat oseckého kláštera dověděl , jaké problémy s nimi jsou , další pobyt zakázal a němečtí hoši se museli i s vychovatelem {bridging, typ=FUNCT, na „hoch“}z kláštera vystěhovat .
>
> (33/In95047_061) Německé chlapce jsme již nezastihli .
> (35/In95047_061) Zastihli jsme tady pouze jediného vychovatele {bridging, typ=FUNCT, na „chlapec“}- pana Fuchse .

Vztah mezi NP „chlapec“ a „vychovatel“ se rozumí jako FUNCT pouze v daném kontextu, kde se vykládá o organizovaných skupinách dětí, které jsou odvezeny do některých měst ČR. Tedy u takové skupiny chlapců vychovatel se předpokládá. Zdá se tedy smysluplné označovat v podobných případech bridging vztah FUNCT.

C.1.2. bridging-vztah *část-celek* (PART)

Vztah části a celku je jeden ze základních bridging vztahů. Vzorové příklady jsou: *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 {žádná koreferenční šipka} a křesly {žádná koreferenční šipka}

!!! 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.

Pozor! Bridging vztah typu PART neoznačujeme pokud jsou uzly propojeny vztahem ACMP. Srov.

> (74 /In94207_76) Na břehu Starnberského jezera u místa utonutí byla postavena kapička s královými daty {žádná koreferenční šipka na „kapička“} narození , vlády a smrti , s křížem {žádná koreferenční šipka na „kapička“} a mramorovou pamětní deskou {žádná koreferenční šipka na „kapička“} .

Příklady:

(44/In9413_006) Dělal jsem bez přestávky celé týdny , často v noci {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“}.

O něco vzdálenější ale přesto označovaný vztah části a celku je v následující větě:

(11/In95047_061) V obci Košťany na Teplicku ještě chlapci ani nebyli , ale místní již dali dohromady petici : " My rodiče dětí základní školy Košťany {bridging, typ=PART, na „Košťany“} protestujeme proti umístění ubytovny pro potrestané německé chlapce .

Vztah bridging_PART označujeme také v případě neodlučitelných části, např. u míst a geografických názvů. Srov. páry *Maroko - marocká města, Maroko – Marrákeš, Německo – Bavorsko – Mnichov* apod. V některých zdrojích je ten vztah vyčleňován zvlášť jako vztah Place/Area (srov. Gardent 2003, odkaz na Clark 1977).

Pozor! (vztah *město - muzeum*)
Vztah mezi např. městem a tím, co v tom městě je, NENÍ bridging_PART, tedy neanotujeme. S podobnými vztahy se v textech setkáváme dost často. Přesně řečeno, muzeum není část města a obraz není část galerie. Srov. např.

(49 /In94207_76) V Mnichově jsou muzea {bridging, typ=PART, na „Mnichov“???} a galerie {žádná koreferenční šipka} se vzácnými obrazy {žádná koreferenční šipka}, částečně (podle času) jsem je navštívil a zhlédl překrásný královský zámek {žádná koreferenční šipka} Nymphenburg .

Podobně jako v předchozím příkladě, v následujícím příkladě stoly a jídelní lístek nejsou přesně řečeno částí pivnice, ale těsně s ní souvisí, NEANOTUJEME. Srov.

(81/In94207_76) Kdykoliv jsem navštívil Mnichov nebo jím projížděl , neopomněl jsem zajít do pověstné mnichovské pivnice Bierbräukeller .
(82/In94207_76) Její obrovská hala , kde se sedělo u dubových stolů {vztah k „pivnice“} a bavorské pivo {vztah k „pivnice“} se nalévalo do litrových korbelů s příklopkami .
(83/In94207_76) Jídelní list {vztah k „pivnice“} byl bohatý na zabíjačkové pochoutky {vztah k „jídelní list“} a masa {vztah k „jídelní list“} , dal se objednat talíř {vztah k „jídelní list“} , kde bylo ode všeho něco , jitrnička {vztah k „jídelní list“} , jelítko {vztah k „jídelní list“} , klobása {vztah k „jídelní list“} a plátek {vztah k „jídelní list“} ovaru nebo větší porce {vztah k „jídelní list“} zvlášť , ke všemu byl čerstvý

rohlík {vztah k „jídelní list“} nebo chléb {vztah k „jídelní list“} a křen {vztah k „jídelní list“} a hořčice {vztah k „jídelní list“} .
(84/In94207_76) S pivem {vztah k „pivnice“} to bylo výtečné .

C.1.3. 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 representativní, 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 subjektovou 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“}.

Pokud nejde pouze o identitu referentů, budeme tyto vztahy označovat.

Srov. také

(18/cmpr9410_028) Proti dřívějšku se však zase objevili noví zájemci o umění z řad podnikatelů {bridging, typ=SET, na „zájemce“}, bank {bridging, typ=SET, na „zájemce“}, spořitelen {bridging, typ=SET, na „zájemce“} a realitních kanceláří {bridging, typ=SET, na „zájemce“}.

(31/cmpr9410_028) Z 1 . pol . 19 . století , kdy se rodila moderní česká krajinomalba , je dnes zájem hlavně o tyto autory : Otec a syn Mánesové {bridging, typ=SET, na „autor“}, Navrátil {bridging, typ=SET, na „autor“}, Piepenhagen {bridging, typ=SET, na „autor“}, Kosárek {bridging, typ=SET, na „autor“}, Bubák {bridging, typ=SET, na „autor“}, Ullík {bridging, typ=SET, na „autor“}, Havránek {bridging, typ=SET, na „autor“}.

V podobných případech věta vypadá přeplněná bledě modrými šipkami…

Sporný příklad 2 (otázka nutnosti a hloubky interpretace).
V následujícím příkladu označíme vztah bridging 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 {bridging, typ=SET na „volontéř“}domluví další hovor.

Sporný příklad 3 (označovaní některých bridging vztahů u nereferenčních NP)
V následujícím příkladě bridging vztah SET neoznačujeme:

(51/In94210_95) Ani pro parlamentní knihovnu nezbude na Malé Straně místo .
(53/In94210_95) Umístit knížky {žádná koreferenční šipka}na speciálně upravených půdách , jako je tomu v Národní knihovně v Klementinu , prý z bezpečnostních důvodů nelze .

NP „knihovna“ v (51) má specifickou referenci, stejně jako v (53), zatímco NP „knížka“ v (53) má generickou referenci, nemůže tedy být elementem „knihovny“ v (51) ani „knihovny“ v (53).

??? 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 nabo PART, na „technika“} blokují 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.

Další příklady:

> (78/In94207_84) I když konzervativní Anglie jeho čin odsoudila , guma se zde chytila a Británie se pro žvýkačku {bridging, typ = PART, na „guma“} stala bránou do Evropy .
> (80/In94207_84) Ještě jeden milník si zaslouží zmínku - zrod bublinové žvýkačky {bridging, typ = PART, na „žvýkačka“}.

> (18/In94207_76) Cestování se značně uvolnilo až do podzimu 1969 , kdy začal být omezen výjezd našich občanů do zahraničí .
> (21/In94207_76) Vzpomínám na takzvané zelené hranice zcela bezbariérové a na dosud nevídanou blahovůli zahraničních a našich celních a policejních orgánů už na jaře 1968 , tedy ještě chvíli před záborem , kdy jsme jeli autem na výlet do Západního Německa {bridging, typ = PART, na „zahraničí“} s Hankou Bělohradskou a Dušanem Hamšíkem .

> (41/cmpr9410_028) Hodnota obrazu však nezávisí jen na autorovi .
> (42/cmpr9410_028) Oleje {bridging, typ = SET, na „obraz“}jednoznačně v čele
> (44/cmpr9410_028) Nejdražší jsou olejomalby {coref_text, typ = NR, na „olej“}, následují tempery {bridging, typ = SET, na „obraz“} a pastely {bridging, typ = SET, na „obraz“}, až o polovinu levnější než olejomalba {coref_text, typ = NR, na „olejmalba“} může být kresba {bridging, typ = SET, na „obraz“} či grafika {bridging, typ = SET, na „obraz“} téhož autora .

(73/In94208_11) Když ho smrt překvapila u psacího stolu , revidoval právě text Prezidentské adresy , kterou pronesl několik dní před tím v Americké ekonomické asociaci .
(74/In94208_11) Poslední věta {bridging, typ = PART, na „text“}, kterou v životě napsal , zněla : Stagnacionisté se mýlí v diagnóze důvodu , proč by kapitalistický proces měl stagnovat .

C.1.4. bridging-vztah sémantického protikladu (CONTRAST)

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

Pozor! Bridging vztah typu CONTRAST neoznačujeme v případě když dané NP jsou potomky uzlu s funktorem ADVS, většinou spojky nebo čárky. Tento funktor vztah kontrastu už vyjadřuje. Srov.

(17/In94208_11) Dočasný podnikatelův zisk bude anulován , ale trvalý zisk {žádná koreferenční šipka} z jeho inovace zůstane zachován společnosti ve formě nižších cen nebo technicky dokonalejších výrobků .

Srov. také neanotujeme CONTRAST u nepřímých potomků uzlu s funktorem ADVS:

Letos by výstavba technického zařízení v sedmi lokalitách stála 120 milionů korun , ale můžeme uvolnit jen 80 milionů {žádná koreferenční šipka} .

Tato skupina se terminologicky prolíná s kontrastem v AČ. Částečně i významově. Jak ukazují příklady, některé bridging-šipky 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 poslancování stalo placenou činností a nároky na jeho vykonávání přiměřeně tomu vzrostly.

V anotaci AČV jeden člen vztahu má c(ontrast), druhý má f(okus):

> (13) Lidi nežvýkají , to jenom krávy {bridging, typ=CONTRAST, na „člověk“}.

Možné jsou i další kombinace.

C.1. 5.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 vnuk {bridging, typ_REST na „#PersPron“} výtečné základy , ač sám vystudoval školu zaměřenou na dopravu.

Pozor! Jak uvedeno výše, bridging šipka 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 vnuk je jednoznačně KZ, přičemž tato KZ je dána pozicí ve větě ale také i sémantickým vztahem děda-vnuk, 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 přikláněla 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.

(105/In94207_84) Všimli jste si někdy , že velká většina skvělých učitelů , kteří ve spojení se žvýkačkou tak rádi mluví o dobytku , vyznává estetiku kouře vycházejícího z úst ?
(106/In94207_84) Krávou tedy člověk být nesmí , drakem {bridging, typ_REST na „kouř“} však ano . . .

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.

C.2. Skupiny bridging vztahů

C.2.1. Vztah „místo – obyvatel“.

Momentálně je zařazen mezi bridging_REST. Srov. např.

(27/In94207_84) Keř , kterého si Kolumbus na ostrově Santo Domingo povšiml , je příbuzným řecké mastiky a jeho mízu místní Indiáni {bridging, typ_REST na „ostrov“} používali stejně jako Řekové .

Podobné vztahy jsou např. *Praha – pražáci, Rusko – Rusové, Mexiko - Mexičan* apod. Srov.

(35/In94207_84) Poté , co byl v roce 1845 jako prezident svržen a na deset let vypovězen na Kubu , vydal se do New Yorku s jedinou myšlenkou - získat zpět vládu nad Mexikem .
(38/In94207_84) Tak jako každý Mexičan {bridging, typ_REST na „Mexiko“}, i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle … .

C.2.2. Vztah typu „autor – kniha“

Pokud není označen funktorem AUTH. Momentálně je zařazen mezi bridging_REST. (není to FUNCT?) Srov. např.

(22/cmpr9410_028) Při výběru obrazu bude hrát určitě velkou roli autor {bridging, typ_REST na „obraz“} .

Ale

(23) Krásná , ale nesignovaná krajinka {žádná koreferenční šipka} neznámého malíře {funktor AUTH, } bude určitě hůře prodejná než slabý Slavíček .

C.2.3. Vztah „věc – majitel“

Neoznačujeme, pokud je zachycen funktorem. Pokud tomu tak není, označujeme jako bridging_REST. Srov.

(54/cmpr9410_028) Obraz výrazně stoupne na ceně , má - li majitel {bridging, typ_REST na „obraz“} doklad o tom , že byl vystaven na výstavě , či je publikován v knize či katalogu .

C.2.4. Vztah mezi stejně vyjádřenými nebo synonymními nekoreferenčními NP

Dost část se potkáváme s páry NP, které jsou stejné nebo synonymní, nejsou koreferenční (třeba jeden člen má generickou, druhý však specifickou referenci) ale které přitom podílejí na koherenci textu. Některé takové vztahy, které jsou nápadné, jsem zařadila do bridging_REST. Srov.

e) stejné NP:

(77/In95047_061) " Sever Čech má za sebou svízelnou minulost , má před sebou po skončení vlády komunistů novou naději , " domnívá se Raimond Strathman , člověk zodpovědný za akci Evangelické diakonie v České republice .
(78/In95047_061) " Staráme se o děti , které se ne vlastním přičiněním dostaly do těžké situace .
(79/In95047_061) Vždyť i ony mají před sebou novou naději {bridging, typ_REST na „naděje“} .

(39/In94207_84) Právě v té době přihrála náhoda Santa Annovi do cesty Thomase Adamse , fotografa a především vynálezce všeho druhu .
(45/In94207_84) Psal se rok 1869 a do hry vstoupila další náhoda {bridging, typ_REST na „náhoda“}.

(27/In94207_84) Keř , kterého si Kolumbus na ostrově Santo Domingo povšiml , je příbuzným řecké mastiky a jeho mízu místní Indiáni používali stejně jako Řekové .
(28/In94207_84) Zatímco karibští Indiáni strčili do úst kousek surové gumy v té podobě , jak jej utrhli od kůry , Mayové na poloostrově Yucatán přivedli žvýkání na vyšší úroveň .
(29/In94207_84) Mízu {bridging, typ_REST na „míza“} stromu sapodilla (achras sapota) sklízeli a upravovali systémem , který se používá dodnes .

??? V posledním příkladě však nejsem si zcela jistá, že se ten vztah má vůbec označovat. Hranice mezi označením a neoznačením je v podobných vztazích dost vágní a má se určovat podle toho, zda daný vztah přispívá ke kohezi textu, totiž podle smyslu.

f) synonymní NP. V následujícím příkladě jsou dvě nekoreferenční NP, které jsou synonymní, mezi sebou zcela jistě sémanticky souvisí a mají vliv na koherenci textu. Obě NP mají tady generickou referenci.. Srov.

(23/In95047_061) Měli časté incidenty s městskou policií : " Díky tomu , že měli dostatek finančních prostředků , byli

```
často opilí , nabalovala se na ně místní mládež , a také si
brali do kláštera holky , " vzpomíná obecní strážník .
(27/In95047_061) I tady si prý chlapci , kteří měli být
vychováváni na faře , užívali děvčat {bridging, typ_REST na „holky“} a
svobody .

(20/In94207_76) Stačilo jen razítko na hranicích , celní
kontroly jejich orgány nedělaly , náš pas - tedy to , že jsme
z Československa , byla sama o sobě průkazná vizitka a vstup
na jejich území ( a stejně tak i na další ) byl hladký .
(23/In94207_76) Naši celníci už nás čekali , viděli nás , jak
vyjíždíme od Němců , zastavili jsme , dali štempl {bridging,
typ_REST na „razítko“} a měli jsme jet dál .
```

C.2.5. Vztah událost – argument

v klasických pracích k bridging (Clark 1977 aj. podle Gardent 2003) je vyčleněn jako zvláštní typ, Srov. např

```
(11/cmpr9410_031.t) Relativně tak stát vynakládá na tržně
konformní podporu malého a středního podnikání přibližně
[1,6]1.6 - [1,8]1.8 % hrubého domácího produktu .
(13/cmpr9410_031.t) V rámci rozpočtové podpory poskytují
ministerstva malým a středním podnikatelům {bridging, typ_REST na
„podnikání“} zvýhodněné informační služby a poradenskou činnost
buď přímo nebo prostřednictvím specializovaných institucí .
(14/cmpr9410_031.t) Potěšitelné ovšem je , že podpora malého a
středního podnikání {bridging, typ_REST na „podnikatel“, coref_text, typ_0 na
„podnikání“} má výrazný regionální aspekt .
```

Na slovesa bych to nedělala. Vztah *podnikat – podnikatel* bych neoznačovala. I když nevím...

Srov. také páry: *spor – účastník konfliktu, léčebna – léčba* apod. Tyto vztahy zatím neoznačujeme, ale nejsem si tím úplně jistá.

C.2.6. Anafora bez koreference

Jsou příklady jednoznačného anaforického odkazu na nekoreferenční entitu. Srov. např „leden – červen“. … „ve stejném období loňského roku“. Ve stejném období – jednoznačná koherence, odkaz na „leden – červen“, ale nemáme nástroj pro její označení. Je tu anafora, ale není koreference. Odkazujeme na místo v kalendáři. Docela typický příklad. Zatím nabízím bridging_REST, protože anotujeme hlavně koreferenci a nemůžeme propojit identitou páry, které nejsou koreferenční.

C.3. Nejednoznačný výběr antecedentů

C.3.1. Spojení se slovy s funkcí „kontejneru“

Ve spojení se slovy s funkcí „kontejneru“ (spousta, řada, milion apod.) bridging šipka vede na kontejner, stejně jako v případě textové koreference (viz B.2.4.4.). Srov.

```
(57/In94207_84) Když o deset let později obrátil ke gumě
pozornost louisvilleský lékárník John Colgan , existovala již
```

řada žvýkačkových milionářů (mezi nimi Adams {bridging, typ_SET na „řada“, nikoliv na „milionář“}) .
(58) Přesto však byly dveře pro zlepšovatele otevřeny dokořán , většina {bridging, typ_SET na „žvýkačkový“} gumy byla stále ještě jen povrchově oslazený či ochucený kousek chicle .

Srov. také

(36/cmpr9410_028) Po dílech uvedených autorů bude nejspíš vždy slušná poptávka .
(37/cmpr9410_028) Průměrné olejomalby {bridging, typ_SET na „dílo“} většiny {bridging, typ_SET na „autor“} z nich {coref_text, typ_0 na „autor“???}stojí kolem 100 tisíc korun .

Křesťané se modlili za usmíření národů...
Více než tisícový zástup {bridging, typ_PART na „křesťan} křesťanů z různých sborů a církví českých zemí a delegace {bridging, typ_PART na „křesťan} křesťanů {coref_text, typ_NR na „křesťan“, funktor APP} z Německa se v sobotu na vrchu Radobýl u Litoměřic modlil za smíření mezi Čechy a sudetskými Němci .

C.3.2. K otázce výběru antecedentu v případě apoziční skupiny:

V případě apoziční skupiny (u koreferujícího člena nebo antecedenta), podobně jako u textové koreference (viz B.2.4.1.), šipku vedeme od na konektor (resp. od něj). Srov.

(21) Vzpomínám na takzvané zelené hranice zcela bezbariérové a na dosud nevídanou blahovůli zahraničních a našich celních a policejních orgánů už na jaře 1968 , tedy ještě chvíli před záborem , kdy jsme jeli autem na výlet do Západního Německa s Hankou Bělohradskou a Dušanem Hamšíkem .
(22) V nějakém městečku za Schirndingem ,#Comma/APPS {bridging, typ_PART na „Německo“} snad v Marktredwitzu, jsem si koupil vynikající umělé květiny , jaké se u nás neviděly , také pětatřiceticentimetrovou opici , huňatou , milou a pár podobných i komických drobností jiných .

C.3.2. K otázce výběru antecedentu v případě koordinační skupiny:

Pokud jde o bridging vztah, šipkou spojujeme především autosémantické uzly, nikoliv spojku (v případě identické koreference je to jinak – viz B.2.4.2..). Srov. např.

(24/cmpr9410_028) Minulé století je bohaté na slavná jména .
(25/cmpr9410_028) Snad vůbec nejvzácnější jsou obrazy Karla Purkyně {bridging, typ_SET na „jméno“} a Jaroslava Čermáka {bridging, typ_SET na „jméno“} .

Je možné však i bridging na spojku, pokud je to ze sémantického hlediska logičtější. Srov.

Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 mil . USD , tj . téměř 2 % HDP .

I když letos a {bridging, typ_CONTRAST na „vloni“} příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu , prognózujeme , že saldo přesto zůstane kladné ve výši 300 - 600 mil . USD ročně .

D. Bridging nebo textová identická – výběr a preference

D.1. K otázce dvojího odkazování (textový a bridging vztahy) a preference:

Pokud NP odkazuje k bližšímu antecedentu bridging vztahem, přičemž je to důležité pro koherenci textu, a zároveň má (třeba i ve vzdálenějším kontextu) textově koreferenční uzel, označujeme oba vztahy – bridging a textový. Srov.

(11/In95047_061)V obci Košťany na Teplicku ještě chlapci ani nebyli , ale místní již dali dohromady petici : " My rodiče dětí základní školy Košťany protestujeme proti umístění ubytovny pro potrestané německé chlapce .
(12/In95047_061) Víme , že na našem území páchali dál trestnou činnost a nevhodně pokřikovali na kolemjdoucí .
(13/In95047_061) Pokud zástupci města neprovedou opatření proti umístění těchto německých chlapců v našem městě {coref_text, typ=SYN na „obec“, bridging, typ=PART, na „dospělý“}, nebudeme posílat svoje děti do školy z důvodu strachu o jejich bezpečnost .

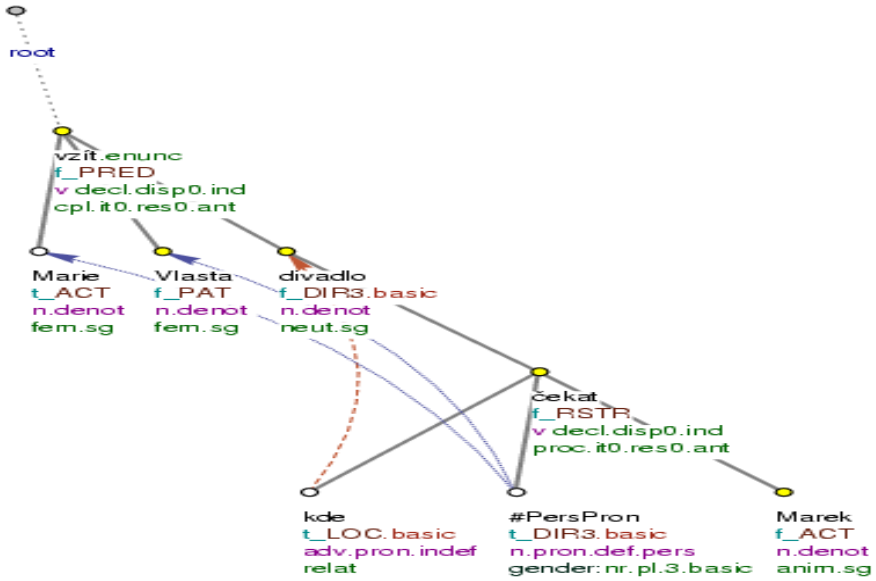
V případě opačného pořadí to neplatí. Srov.

(2/In94207_84) Vybrané kapitoly z dějin žvýkačky.
(9/In94207_84) Jeho milovaný kousek žvýkací gumy { bridging, typ=REST, na „žvýkačka“}, který si tak pečlivě odložil na spodek desky stolu , se stal kořistí nepřítele a asi jej čeká potupný konec v odpadním koši .
(26/In94207_84) Pro historii žvýkací gumy {coref_text, typ=NR na „žvýkačka“}, jak ji známe dnes , se však musíme přenést na jiný kontinent .

NP „žvýkací guma“ v (26) odkazuje k NP „žvýkačka“ ve větě (2). I když mezi (2) a (26) je věta (9) se stejnou NP „žvýkací guma“, ta není koreferenční se „žvýkací gumou“ v (26).

D.2.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 budeme podobné trojice propojovala jako bridging, typ_SET.

Pozor! Pokud však antecedenty takového vztahu jsou (technicky) přímé potomky spojky, vedeme textovou tmavomodrou šipku na tuto spojku. Srov.

> (42 /In94207_84) Asi rok se Adams a jeho nejstarší syn snažili - chicle vařili , čistili , přidávali množství různých látek a míchali s pravým kaučukem .
> (44 /In94207_84) Když asi po roce #PersPron {coref_text, typ=0 na #Comma, **nikoliv** bridging, typ=SET na Adams ani syn}své úsilí vzdali , rozhodl se Adams , že vše , co mu z chicle ještě zbylo , hodí do řeky .

??? U toho řešení si nejsem jistá. Teď to ale tak všude je – vždyť je to #PersPron, tedy patří do základní anotace textové koreference.

D.3. K otázce výběru antecedentů u několika sémantický spojených řetězců se specifickou referencí (textová + bridging):

D.3.1. Dlouhé mezi sebou propojené řetězce

Situace, když entity se specifickou referencí se jednou vyskytují jako jednotlivé jednotky, jednou jako část celku a pak zase jednotlivě. V „parlamentním textu“ (ln94210_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]búra 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...

Názornější příklad je následující:

> (67/In95047_061) Pavel Vondráček : " Termín převýchova znám pouze z nacistického [slovníku] a komunistického slovníku .
> (68/In95047_061) Na převýchovu se pokud vím , posílali ti , kteří měli podle těchto zrůdných režimů nevhodný původ .
> (69/In95047_061) Židé , cikáni , šlechta , podnikatelé , kulaci a jiní .

Máme tady dvě možnosti:

e) odkázat „režim“ textovou identickou koreferencí na spojku „a“
f) odkázat „režim“ bridging vztahem typu PART zvlášť na „nacistický slovník“ a „komunistický slovník“

Situaci ještě komplikuje to, že slovník ≠ režim a v podstatě bychom měli referovat na adjektiva nikoliv na substantiva, ale to už jsou drobnosti a nekonsekventnosti autorského stylu.
Obě možnosti mají stejné právo na existenci. Co vybrat – vůbec nevím. Momentálně jsou obě zachyceny.

Stejné argumenty platí pro pár *„ti , kteří“ - Židé , cikáni , šlechta , podnikatelé , kulaci a jiní*. Pravděpodobné jednodušším řešením bude označovvat v takovýh případech jenom textovou koreferenci se spojkou, abychom předešli velkému množství bledě modrých bridging šipek. Ale je to slabý argument.

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

D.3.2. „faktory – jeden z faktorů“

Situaci, když v antecedentu je NP „*x*“, druhý člen páru je konstrukce „jeden (některé) z *x*“ řešíme jako v následujícím příkladu:

(16/In94208_11)V praxi se tato rovnováha realizuje tím , že se každý faktor[14] chová [rutinnímu]rutinním a adaptivním způsobem .
(17/In94208_11)Dynamika kapitalistického systému vzniká teprve tím , že se jeden {bridging, typ_SET na „faktor“ v **(16)**} z faktorů {coref_text, typ=0 na „faktor“} začne chovat netradičně .

Další možné řešení je odkaz u jeden dopředu na „faktor v téže větě“:

(16/In94208_11)V praxi se tato rovnováha realizuje tím , že se každý faktor chová [rutinnímu]rutinním a adaptivním způsobem .
(17/In94208_11)Dynamika kapitalistického systému vzniká teprve tím , že se jeden {bridging, typ_SET na „faktor“ v **(17)**} z faktorů {coref_text, typ=0 na „faktor“} začne chovat netradičně .

Podobně se chovají generické NP ve spojením s „konteinery“. V párech jako *sklenice vína, krabice gumy* konteier může mít v kontextu libovolný typ reference, zatímco syntakticky závislý uzel bude mít vždy generickou interpretaci. Propojovat bychom měli oba řetzce – jak konteinerů tak i to co obsahují.

Srov. příklad:

(73/In94207_84) V rámci reklamní kampaně předal osobně každému členu washingtonského Kongresu jednu krabici {specifická reference, distributivní} své gumy {generická reference} značky Yucatan .
(76/In94207_84) I král dostal svou krabici {specifická reference, žádná šipka} gumy {generická reference, coref_text, typ=NR na „guma“ v (73)} a nádavkem i doslova trhoveckou prezentaci .

[14] V daném případě sice jde o distributivní referenci, ale v daném kontextu je to totéž co „všechny faktory“, takže považujeme to za množinu.

(77/In94207_84) Samozřejmě , že novinové zprávy o králi s krabicí {specifická reference, coref_text, typ=0 na „krabice“ v (76)} gumy {generická reference, coref_text, typ=NR na „guma“ v (76)} byly reklamou k nezaplacení .
(78/In94207_84) I když konzervativní Anglie jeho čin odsoudila , guma {generická reference, coref_text, typ=NR na „guma“ v (77)} se zde chytila a Británie se pro žvýkačku stala bránou do Evropy .

Případ „zaměstnanci – každý ze zaměstnanců“

Poněvadž „každý“ v tektogramatickém stromě má substantivní platnost, v konstrukci „každý ze zaměstnanců“ anotujeme koreferenci od něj, PP „ze zaměstnanců“ necháváme bez šipky, jako závislý uzel. Srov.

(13)Podle přesvědčení majitelů dosáhla prosperity zejména proto , že zaměstnává lidi , na které { coref_gram, na „člověk“} se může spolehnout .
(14) Kritéria výběru jsou přísná .
(15) Každý { coref_text, typ=NR na „který“} ze zaměstnanců musí být odborníkem .

??? nějaké pravidlo, které by vyhledávalo závislé uzly s DIR1? Nebo udělat to nějak jinak?

* * *

Připustitelná nepřesnost v určování bridging vztahů

Některé vztahy vypadají velice výrazně jako jeden typ bridging, uzly mají mezi sebou stejnou sémantickou souvislost jako v těch vztazích, ale konkrétní NP ve větě tam přesně nezapadají, např. slovnědruhově. Srov.

(83/In95047_061) V Evangelické vesnici mládeže se dodržuje zásada , že každý vychovatel vedle své pedagogické odbornosti je vyučen nějakému řemeslu .
(84/In95047_061) Mezi vychovateli jsou obchodníci {bridging, typ_SET na „remeslo“}, malíři {bridging, typ_SET na „remeslo“} pokojů , kuchaři {bridging, typ_SET na „remeslo“}, truhláři {bridging, typ_SET na „remeslo“} .

Přesně řečeno „obchodník“ , „malíř“, „kuchař“ a „truhlář“ nejsou podmnožinou „řemeslo“, ale je to v podstatě zvláštnost syntaktické a stylistické struktury textu, budeme to tedy anotovat jako bridging typu SET.

Srov. také

(52 /In94207_76) Byl romantik , velký ctitel Richarda Wagnera , který si dal postavit velmi nákladný zámek Neuschwanstein neméně romantický na vrchu a skále s věžemi a zařídil jej s největším přepychem a s patinou tajemnosti .
(65 /In94207_76) Obdiv {bridging, typ_REST na „ctitel“}Richarda Wagnera a jeho děl z německé mytologie uznal u Ludvíka za normální projev jeho vkusu .

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 vyhledávaní hypertémat článků. Z těchto důvodů hypertématické vztahy ve stávající anotaci neoznačujeme.

D. Speciální typy reference (coref_special)

Exoforické odkazy

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 patři především časová (*v tomto roce*) ale i prostorová deixe. Srov. příklady:

> `(37/In94207_84) Jednou z nejžádanějších komodit na světě byl v_té době {coref_special, typ exoph nebo segm???}kaučuk , kterého nebylo dost a dovážel se z daleka .`

S odkazy typu „v té době“ mám problém už dávno. Je to v podstatě endoforické odkazování, odkazuje se na čas, o kterém se mluví v textu, čili je tam dokonce jakási skrytá korelace. Pokud bychom ten vztah označili jak `exoph`, odkazoval by na současnost, což není pravda. Avšak na druhé straně nedá se přesně určit, na který segment textu ten odkaz vede, čili `segm` to také nebude. Ještě jedna možnost označovat to jako textovou koreferenci na řídící uzel předchozího stromu, pokud časový kontext je vyjádřen jednou větou a `segm`, pokud je takových vět několik. Ale takový předchozí strom ne vždy něco tvrdí o času, resp. informace o času je jenom část podávané v této větě informace, čili to by už nebyla identická koreference, navíc tam je přítomen ještě jeden myšlenkový krok (v době, kdy se to všechno dělo). …
Momentálně jsem takové případy označovala jako `segm`. Srov. také:

(38/In94207_84) Tak jako každý Mexičan , i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle (prý z mayského slova tsictle) , a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku .
(39/In94207_84) Právě v té době {coref_special, typ exoph nebo segm???, nebo coref_text, typ_0 na „a“} přihrála náhoda Santa Annovi do cesty Thomase Adamse , fotografa a především vynálezce všeho druhu .

Srov. také příklady z SYN2005:

Po chvilce 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)

Odkazy typu exoph dodáváme pouze v případě opravdové euforické deixe (ukázaní prstem), nikoliv jakéhokoli mimojazykového odkazu. Např. označujeme v dialogu "tahle budova je Šternberský palác", nikoliv však konstrukce typu "příští rok", "v současné době" apod.

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}.

Další exoph odkazy jsem dodávala u adverbií – pokud u nich anotujeme rozšířenou textovou koreferenci, měli bychom anotovat i speciální koreferenci. Srov. např.

(104/In94207_84) A tu {coref_special, typ exoph} se dostáváme zpět k počátku tohoto textu.

Pozor! Exoforickou šipku neanotujeme v případě když deiktický výraz je součástí lexikální sémantky daného slova (výrazy typu *dnes, zítra, letos* apod.) Kdybychom to potřebovali, dá se to vždy dodělat automaticky.

Exoforický odkaz nezaznamenáváme:

Angel říká , že fronty se každým dnem znatelně prodlužují . " Viděl jsem šňupat opravdový dámy , jsou tu i lidi , který vypadaj , jako by umírali na AIDS .
Je hrozný , jak jim takovýhle život užírá rozumný myšlení rychleji než blesk . "

Odkazy na segmenty textu

Odkazy na segmenty textu dodáváme v případech, kdy jmenná fráze (většinou s identifikátorem) odkazuje na více než jednu větu v předchozím kontextu. Srov. např.

(10/cmpr9410_001.t)Celní unie bude sice existovat na papíře ještě dalších dvanáct měsíců (a třeba i déle) , ale v praxi dostanou vzájemné vztahy punc tvrdosti mezinárodního obchodu .
(11/cmpr9410_001.t)Poroste administrativa
(10/cmpr9410_001.t)Jistotu v tomto směru {coref_special, typ segm} dávají nejnovější kroky vlády SR , která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční [provinencie]provenience .

Srov. také

V článku jsme odpovídali na dotaz naší pardubické čtenářky , kde by měla uzavřít životní pojištění , aby platila co nejméně a získala co nejvíce . Při výběru pojišťovny jsme zvažovali , kolik by musela zaplatit ročně na pojistném , zda by se mohla připojistit na úraz , zda by byla okamžitě po uzavření pojistné smlouvy pojištěna na sjednanou pojistnou částku a konečně zda si bude moci v případě náhlé potřeby vypůjčit větší sumu peněz z dosud zaplaceného pojistného aniž by to mělo vliv na výši pojistné částky .
(6 vět)
V uvedeném příkladu {coref_special, typ segm} by byla minimálně 30 % a maximálně 50 % .

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.

Občas můžeme potřebovat, nikoliv realizovat segm odkaz dopředu. Neoznačujeme to nijak, jenom jako ukázka, že to také existuje:

Do redakce nám přišly ohlasy (tady by se hodil segm dopředu).
Z dopisu Bedřicha Kováře , ředitele úseku pojištění osob v České pojišťovně vyjímáme :
Podíl na zisku je v článku zmíněn , avšak v případě pojištění nabízeného Českou pojišťovnou je důležitou skutečností , že

pojištěný má mimo sjednanou pojistnou částku zaručenou zvláštní prémii a navíc valorizaci . *atd*

Preference
Pokud můžeme spolehlivě odkázat typem SET, preferujeme SET, nikoliv segm.

- Stála vás kvalita hodně peněz a potu ?
- Museli jsme se přizpůsobit tržní filozofii .
Dříve jsme měli za úkol jen nasytit trh množstvím výrobků a na jakost se nehledělo .
Nyní jsou požadavky opačné .
Proto jsme zpřísnili vlastní kontrolu .
Inovovali jsme také receptury pracích prášků , zvýšili podíl účinných látek a parfémů .
U detergentu Toto jsme například řešili problém s udržením stálé kvality , protože jednotlivé partie byly nevyvážené .
Investovali jsme dva miliony korun do nákupu pásových vah , zpřesnili dávkování a jakost pracího prášku stabilizovali .
– V těchto opatřeních {bridging, typ SUB_SET na *zpřísnit, inovovat, zvýšit, řešit, investovat, zpřesnit, stabilizovat* } vidíte podstatu komerčního úspěchu ?

–

Jako coref_special, typ_segm řešíme také případy odkazy na segmenty uvnitř jednoho stromu, kdy technicky nejde odkázat na daný segmant. Srov. dále v textu v Technických problémech.

* * *

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 použit 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 zůstává přitom identická koreference.

Koreference u spojení obecného jména a pojmenované entity

Při anotaci koreference NP s obecným jménem a pojmenovanou entitou s funktorem ID nebo RSTR (*firma Struktura, země Španelsko, v Sekaninově ulici, projekt Světlo v temotách* apod.) vzniká otázka, který uzel má koreferovat – pojmenovaná entita nebo obecné jméno. Z hlediska struktury stromu a pravidel anotace jiných úseků je logické koreferovat na formálně řídící uzel (tj. na *firma, země, ulice, projekt* apod.), ale zkazí nám to sémantickou návaznost řetězců. Avšak nedá se nic dělat, koreferujeme vždy na řídící uzel. Srov.

(15/In95047_061) V Košťanech totiž zakoupila dům firma *Struktura* , která se u nás rozmísťováním německých chlapců zabývá .
(28/In95047_061) Posledním místem , kam byli chlapci firmou {coref_text, typ_0 na „který“} *Struktura* umístěni , byl bývalý dům dětí a mládeže v Duchcově .
(48/In95047_061) Tajemná Struktura {coref_text, typ_0 na „firma“}
(49/In95047_061) Ten , kdo ve skutečnosti německé chlapce v severočeském pohraničí umísťoval [], byla firma {coref_text, typ_0 na „Struktura“} *Struktura s . r . o .* [], která se zabývá sociálním managementem .
(50/In95047_061) Její {coref_text, typ_0 na „firma“} zástupce ing . Šedivý však veškerou odpovědnost za krizovou situaci odmítá .
(52/In95047_061) Mezitím starosta obce Košťany Jindřich Abrhám požádal o pomoc člena branně - bezpečnostního výboru parlamentu poslance Čapka (Levý blok) , který se obrátil na ministra Rumla s žádostí o prošetření činnosti firmy {coref_text, typ_0 na #PersPron}*Struktura* .

??? V případě **páru dvou obecných jmen** je situace složitější. Srov.

(67/In95047_061) Pavel Vondráček : " Termín {PAT} převýchova {ID} znám pouze z nacistického a komunistického slovníku .
(68/In95047_061) Na převýchovu {coref_text, typ_0 na „termín“ nebo na „převýchova“} se pokud vím , posílali ti , kteří měli podle těchto zrůdných režimů nevhodný původ .

Momentálně šipka vede k NP „termín“, ale přesně řečeno tam není identická koreference.

Je to však logické z hlediska teorie reference. Přesně řečeno, jména s funktorem ID nejsou plnohodnotně referenční, referují v podstatě samy na sebe, autonymně. Takže vybrané řešení má také referenční oprávnění. Podobně se takové situace řeší v Chiarcos,Krasavina2005:29. Ale mají to jednodušší, protože to řeší na složkách.

Anotace částí pojmenovaných entit

Části pojmenovaných entit anotujeme podle smyslu - pokud část NE ma nějakou slušnou referenci, resp. v názvu je něco, co potom přirozeně vstupuje do samostatného referenčního řetězce, tak to označíme. Pokud ne, raději necháme nepropojené. Srov. např.

České Budějovice - České Budějovice	spojíme pouze *Budějovice,* nikoliv *české*

Nbpříklad – dodat další příklady

Možnosti automatizace rozšířené anotace koreference a bridging vztahů

- Named entities – kombinací technik by to snad šlo. Často se opakují plnohodnotné názvy, nemění se tolik na synonyma, vystupují v subjektní pozicí s dalším #PersPron, lze pak řešit morfologicky, u názvů mist nejčastější odkaz je adverbium, např. *Německo - tam*. Je důležité nejenom pro kohezi textu ale i pro řešení aktuálních úkolů počitačové lingvistiky.
- v řetězci s named entity ta named entita bude zpravidla na začátku NB!!!

Co má být doděláno automaticky:

CPR a kontrast
patří do první skupiny patří … - nedělala jsem tam šipky bridging PART. Dodělat automaticky?
koreference subjektu a jmenné části přísudku.

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-anafory, kromě vyhledávání správných párů, je problematické (ne)zařazení do typu REST – existence nejasného sémantického vztahu. Existuje dost pohraničních příkladů mezi SET a PART v případě nepočitatelných a abstraktních referentů.

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

Úroveň rozšířené koreference je textová, představuje již v podstatě nadstavbu nad tektogramatickou rovinou. Anotace ve mnoha případech záleží na chápání textu, které může být u různých anotatorů odlišné. Ve srovnání s původní pronominální koreferencí, víceznačnost vztahů v rozšířené identické koreferenci a hlavně v bridging vztazích je mnohem větší. Čím dále se vzdalujeme od fonologií směrem k textu, tím je polysémie znaků větší (Srov. Melčuk 1974) Srov. příklady

Šedesát tři vězňů , kteří vykonávají trest odnětí svobody v České republice , požádalo za první půlrok o předání do věznic na území Slovenska .
Informoval o tom včera tiskový mluvčí generálního ředitelství Vězeňské služby ČR Eduard Vacek .
Dodal , že loni podalo tuto žádost 200 odsouzených .
Praktické předávání však začalo až letos v červnu , kdy bylo předáno 16 odsouzených .
Další dva budou převezeni počátkem září .
Malý počet předaných osob je podle něj způsoben především administrativními problémy .

V poslední větě jsou možné dvě interpretace. NP „malý počet“ můžeme odkázat identickou koreferencí na „16 odsouzených“ a tím „malý počet“ chápeme jako množinu odsouzených, kteří již byli předáni. Druhá možnost je odkázat „malý počet“ pomocí bridging vztahu typu SET na „16 odsouzených“ a „další dva“. Tím do „počtu“ přidáváme také ty, kteří budou brzy předáni.

Také v případě dávat vs. nedávat šipku se mohou anotatoři spravedlivě rozcházet. Srov.

Na dotaz , zda si ODS zaplatí předvolební reklamu v televizi , řekla tisková mluvčí J . Petrová , že zákon o televizním a rozhlasovém vysílání to neumožňuje.

Zda si můžeme dovolit propojit bridging šipkou typu PART NP „televize“ a „televizní a rozhlasové vysílání“ je na anotátorovi. Chápe-li anotátor v daném kontextu „zákon o televizním a rozhlasovém vysílání“ jako zákon o televizi a rozhlasu, nikoliv o procesu, může tyto uzly propojit. Není to však jednoznačně správné řešení.

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.

Technické problémy kombinace rozšířené a původní anotace koreference.

5. Apozice, na kterou vede šipka gramatické koreference a rekonstruované uzly nově propojené textovou koreferencí. Mely by být tyto uzly spojené, ale technicky to nevychází. Srov. násl. příklad a obr.1

(4/cmpr9410_001.t) Až na jednu jedinou , tu hlavní , která proklamativně zakotvuje existenci Česko - Slovenské celní unie , a která má výpovědní lhůtu jeden rok .

2. Potřebovali bychom odkázat na podstrom, který není oddělitelný. Řešíme to v anotaci zatím jako coref_special, typ_segm. Srov. násl. příklad a obr.2

(79/In94208_11.t) Na Harvardu , kde tento třešťský rodák učil od raných třicátých let až do své smrti , kolovala autobiografická poznámka : Chtěl jsem být největším ekonomem na světě , největším milencem na světě a největším jezdcem na světě .
(80/In94208_11.t) Vzhledem k pokročilému věku třetí cíl už nestihnu .

NP „třetí cíl“ by měla správně odkázat na podstrom „být největším jezdcem na světě“, ale nejde to, protože „být“ je v předchozí větě jenom jednou.

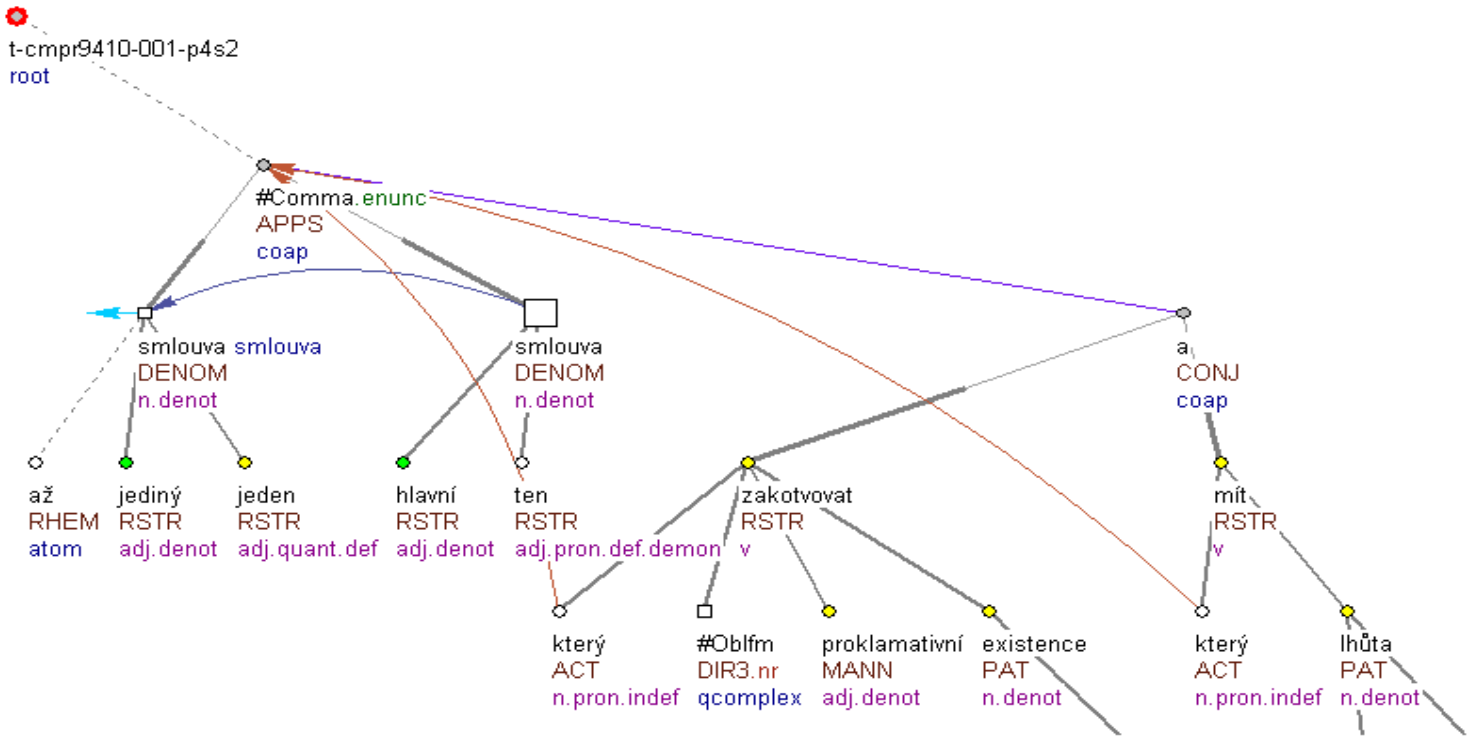
Srov. podobně v

(4/In94206_38.t) Šedesát tři vězňů , kteří vykonávají trest odnětí svobody v České republice , požádalo za první půlrok o předání do věznic na území Slovenska .
(5/In94206_38.t) Informoval o tom včera tiskový mluvčí generálního ředitelství Vězeňské služby ČR Eduard Vacek .
(6/In94206_38.t) Dodal , že loni podalo tuto žádost 200 odsouzených .

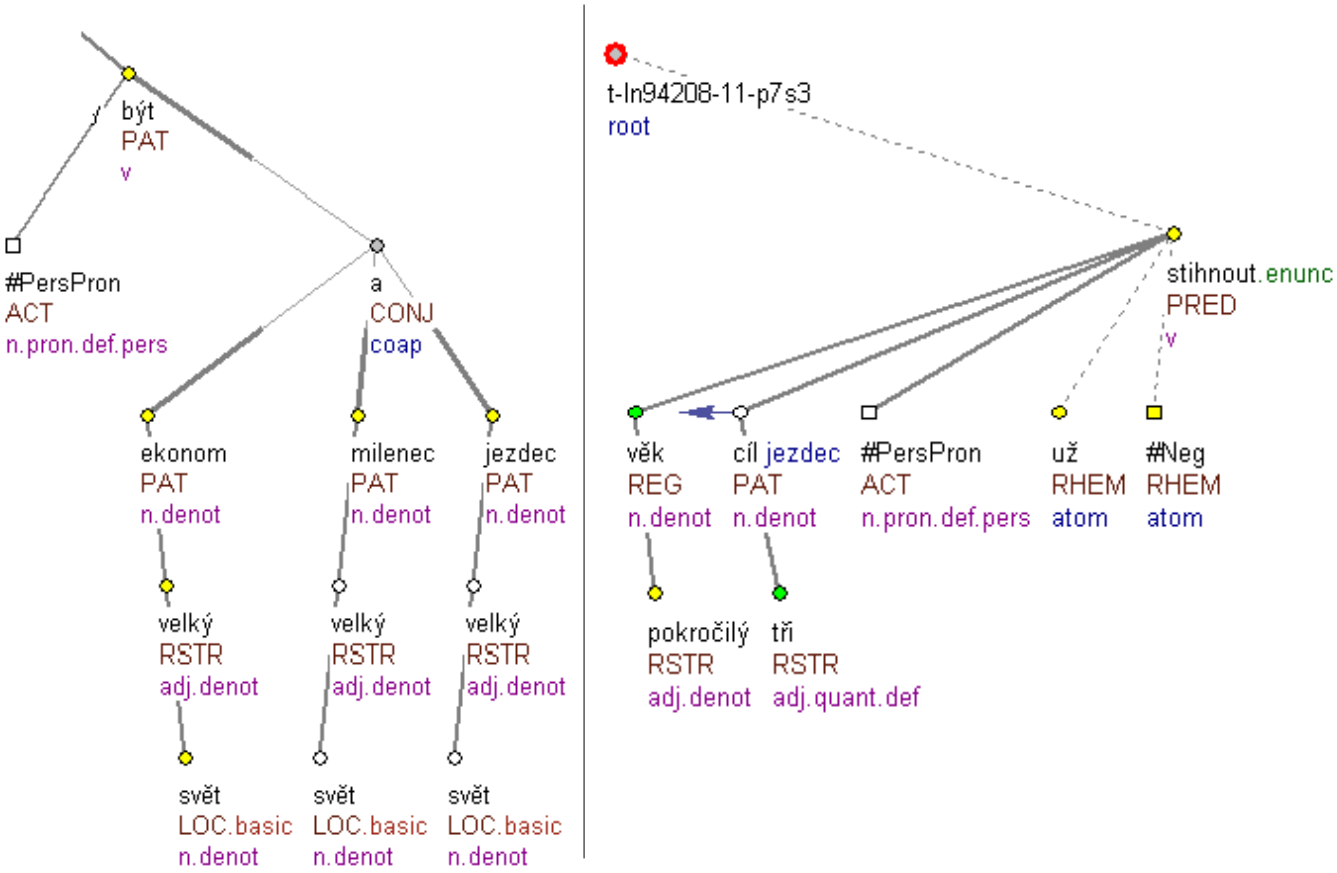
3. Podobně v (20/cmpr9410_001.t) - struktura stromu v (19/cmpr9410_001.t) nedovolí odkázat NP „tato funkce“ na antecedentní podstrom bez subjektu. Řešíme to v anotaci zatím jako coref_special, typ_segm. Srov. násl. příklad a obr.3

(19/cmpr9410_001.t) Od 1 . dubna nebude ÚNMS SR rozhodnutí české zkušebny potvrzovat .
(20/cmpr9410_001.t) Tato funkce přejde na příslušnou slovenskou zkušebnu (SPPI - Slovenskou potravinářskou a zemědělskou inspekci) , která bude vydávat na základě dodaných podkladů příslušné certifikáty .

Obr. 1

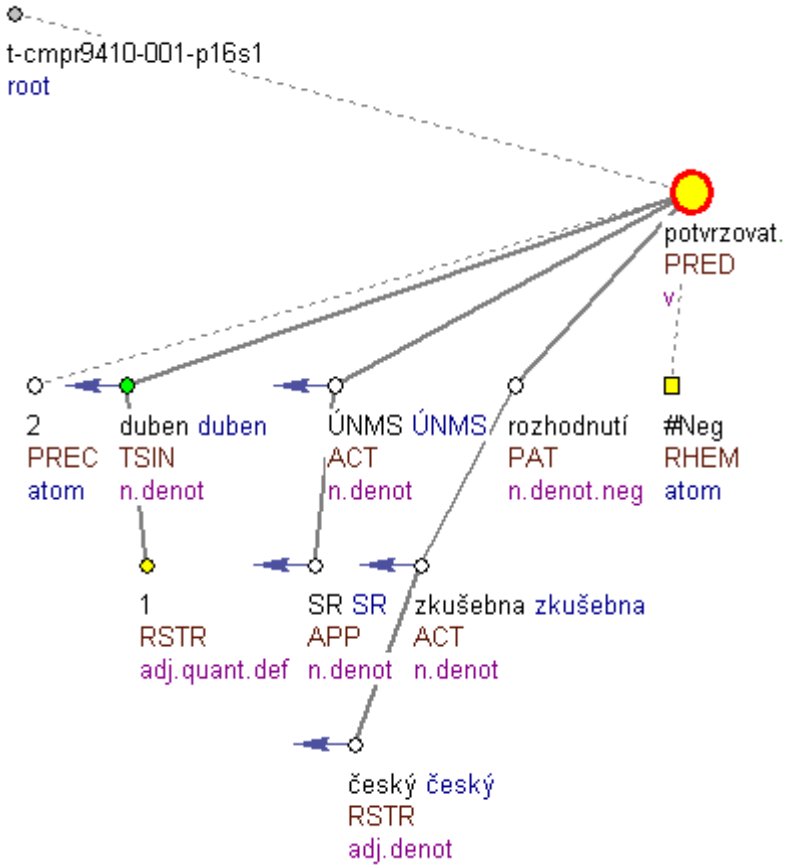

t-cmpr9410-001-p4s2
root
#Comma.enunc
APPS
coap
smlouva smlouva
DENOM
n.denot
smlouva
DENOM
n.denot
a
CONJ
coap
až
RHEM
atom
jediný
RSTR
adj.denot
jeden
RSTR
adj.quant.def
hlavní
RSTR
adj.denot
ten
RSTR
adj.pron.def.demon
zakotvovat
RSTR
v
mít
RSTR
v
který
ACT
n.pron.indef
#Oblfm
DIR3.nr
qcomplex
proklamativní
MANN
adj.denot
existence
PAT
n.denot
který
ACT
n.pron.indef
lhůta
PAT
n.denot

Obr.2

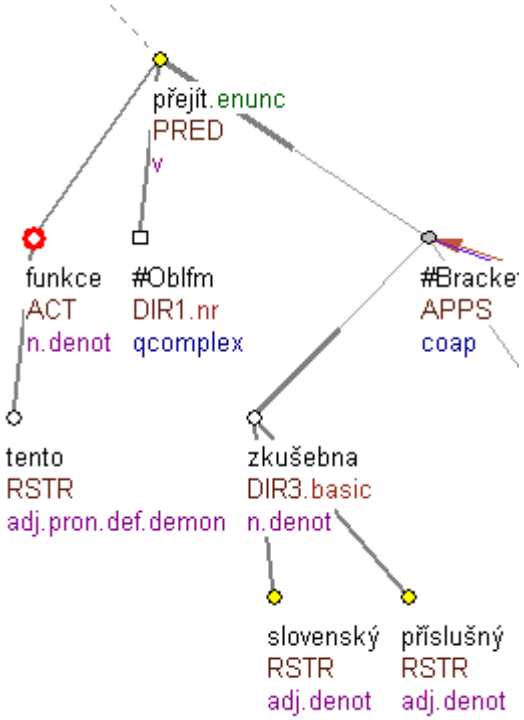


(79/In94208_11.t) (80/In94208_11.t)

Obr. 3
(20/cmpr9410_001.t)



(20/cmpr9410_001.t)



Příloha X: SROVNÁNÍ KONCEPCE PDT A JINÝCH PŘÍSTUPŮ

Chiarcos, Krasavina 2005

PŘEVÉST DO TABULKY

Koncepce anotace v Chiarcos, Krasavina 2005 je velice zajímavá z různých hledisek. Zmíním několik principů, které jsou srovnatelné s našimi, nebo se v něčem liší, nebo mě jinak nějak zaujali.

- při anotaci koreference jak v základní, tak v rozšířené verzi, nikdy neberou v úvahu adjektiva. Anotují se hlavně jmenné a předložkové fráze. Koreferenci na propozice je jenom v rozšířené verzi, zatímco my to máme už v původní anotaci textové koreference (pokud na propozici odkazuje zájmeno)
- mají seznam pronominálních adverbií, které se anotují. Mohlo by se nám to také hodit pro předanotaci
- NP/PP v koordinačních konstrukcích jsou anotovány dvakrát – jako celek a zvlášť elementy. V naši anotaci se anotuje jenom jednou, hlavně na spojku, aspoň v případě textové koreference. U bridging se to řeší často podle smyslu a intuice. Teoreticky nevidím problém anotovat od spojky identickou anaforu a od jednotlivých uzlů pokud nutno bridging.
- v projektu se autoři snaží anotovat anaforické vztahy, nikoliv řešit jakoukoli koreferenci. Takové omezení samozřejmě zjednodušuje a upřesňuje výběr „markables“ a řešení koreference, avšak my si to nemůžeme dovolit, protože těžko ve slovanském jazyce, který nemá gramatickou kategorii členu, vyčleníme určité deskripce.
- preference identity před bridging, i když bridging antecedent je blíž v textu, než antecedent identický – máme to stejně
- proces anotace probíhá jiným směrem, než u nás. V Chiarcos, Krasavina 2005 se nejprve anotují „markables“ jako potenciální jednotky spojené koreferencí, a pak se pro ně hlédají případné antecedenty. V našem projektu vyhledáváme rovou koreferenční páry, přip. pro pravý konec řetězce hledáme další koreferující uzel.
- gramatická koreference, která se u nás řeší zvlášť, je v Chiarcos, Krasavina 2005 rozdělena na základní a rozšířenou, třeba reflexivní zájmena se řeší až v rozšířené verzi, tj. nejsou myšlena pro budoucí automatické zpracování. To máme lepší :-)
- pokud vztah přesahuje větu, může být v Chiarcos, Krasavina 2005 jenom anaforický, tj. šipka vede vždy dozadu.
- zvlášť je zaveden systém typů pro nejasné případy. Např. v případě nejednoznačného výběru antecedenta, anotátor označí daný příklad atributem ambig_ante, a na tento uzel by neměla vést žádná koreferenční šipka (aby se předešlo zmatkům). Přijde mi to přínosné, protože pak „čisté“ případy jsou vyznačeny a všechny evaluace se mohou provádět jenom na nich.
- anotační proces zahrnuje řešení velkého počtu atributů. Je to pravděpodobně časově o moc náročnější anotace, než anotace koreference na PDT. Anotátor označuje nejenom typ koreference, ale také atributy „direct speech“, typ fráze (NP, PP, jiná), typ NP (named entity, určitá NP, neurčitá NP, osobní zájmeno apod.) a atribut „typ ambiguity“, který má až 7 významů. Většina těchto informací je zahrnuta v tektogramatických atributech PDT 2.0. Pozitivní na anotaci v Chiarcos, Krasavina 2005 mi však přijde to, že každý atribut má význam „other“, kam anotátor může zahodit nejasné příklady. Na druhé straně je to však dost nebezpečné – aby nakonec všechno se nedostalo do odpadkového koše.
- V atributech na rozšířené koreferenci mají informaci o sémantické třídě (abstraktní, osoba, materiální objekt, událost apod.) Něco derivují z WordNetu. Dost nám to chybí...
- při anotaci predikativních NP se nevyčleňují identické konstrukce. My také ne.

- v rozšířené anotaci koreference jednotky nemohou sloužit antecedentem pro anaforické vztahy, aby se dalo propojit základní a rozšířenou anotaci. Tím se však „zbaví“ koreferenčních řetězců.
- v rozšířené anotaci koreference jako antecedent slouží význam celého souborů předchozích antecedentů, informace se jakoby množí a doplňuje se. Pro určení typu vztahu se díváme na celý předchozí kontext, nikoliv jenom na poslední antecedent.
- bridging se neanotuje, pokud jakékoli části daných NP jsou spojené identickou koreferencí. Takže bridging se chápe jako dost sekundární – anotujeme pouze v případě když to visí úplně nepropojené.
- bridging se neanotuje, pokud NP jsou významově spojené nikoliv kontextem, ale obecnou znalostí světa. To je dost vágní definice, jako kontext je v Chiarcos, Krasavina 2005 chápána i slovníková informace. Spíše neanotují jenom okazionální implicitní vztahy.
- bridging se anotuje jenom u slov s plnohodnotnou lexikální sémantikou – žádná zájmena a elipsy. To mi přijde dost logické – také se k tomu přikláním
- bridging se neanotuje přes uvozovky – jenom uvnitř přímé řeči a nepřímé řeči. Taky mi to přijde dost logické.

References:

Carlson, L.; Marcu, D. and Okurowski, M. (2003): Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory, in Current Directions in Discourse and Dialogue, pp. 85-112, Jan van Kuppevelt and Ronnie Smith eds., Kluwer Academic Publishers

Claire, Gardent, Helene Manuelian, and Eric Kow. 2003. Which bridges for bridging definite descriptions? In *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, pages 69-76, Budapest.

Clark, Herbert H.: Bridging, Proceedings of the 1975 workshop on Theoretical issues in natural language processing, June 10-13, 1975, Cambridge, Massachusetts

Frege, G. Über Begriff und Gegenstand, Vierteljahresschrift fü r wissenschaftliche Philosophie. Leipzig, Vol. 16, 1892, s. 192-205

Hirschman, L.: 1998. MUC-7 coreference task definition version 3.0. In N. Chinchor, editor, In Proc. of the 7th Message Understanding Conference.

Kučová, L. a kol. Anotování koreference v Pražském závislostním korpusu. (2003)

Lezin G.V.: ON AUTOMATIC DISCLOSURE OF REFERENCIAL COHERENCY IN NARRATIVE TEXT// In Proceedings to 28th Annual Meeting „DIALOG-2007“, June 4-8, 2007, Bekasovo, Rusko

Mengel, A. a kol.: MATE Dialogue Annotation Guidelines,
8 January 2000, Mengel, A., Dybkjaer, L., Garrido, J.M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C.

Mikulová, M. a kol. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka (2005)

Miller, G. et al. 1993. Five papers in WordNet. Technical Report CSL Report ~3, Cognitive Science Laboratory, Princeton University.

Nedoluzhko: UZ ten a generické NP v češtině, 2003

Poesio, M. 2004. "The MATE/GNOME Scheme for Anaphoric Annotation, Revisited", Proc. of SIGDIAL, Boston, April.

Poesio, M. 2000. Coreference. In: MATE Dialogue Annotation Guidelines,
8 January 2000, Mengel, A., Dybkjaer, L., Garrido, J.M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C.

Stede M. *2004*. *The Potsdam Commentary Corpus*. In Proceedings of the ACL. Workshop on Discourse Annotation, Barcelona, *2004*. 96-102.

Vieira, Renata and Simone Teufel : Towards resolution of bridging descriptions.
In proceedings to 35th Annual Meeting of the Association for Computational Linguistics

Мельчук,И. А. Опыт теории лингвистических моделей «Смысл ↔ Текст». М., 1974 (2-е изд., 1999).

Падучева, Е.В. О референции языковых выражений с непредметным значением. НТИ, сер.2, N 1, 1986.

Степанов, Ю.С.: Имена, предикаты, предложения (семиологическая грамматика). Москва, Едиториал УРСС, 2004