Anotace jmenné koreference a asociační anafory na tektogramatické rovině

Reference - Koreference

- *Helena poprosila maminku, aby na ni počkala*.

antecedent

koreferující člen

Koreference v PDT 2.0

Na tektogramatické rovině se anotuje:

- **gramatická koreference**
- **textová koreference**
  - zájmenná
  - jmenná
- **bridging anafora**

HOTOVO

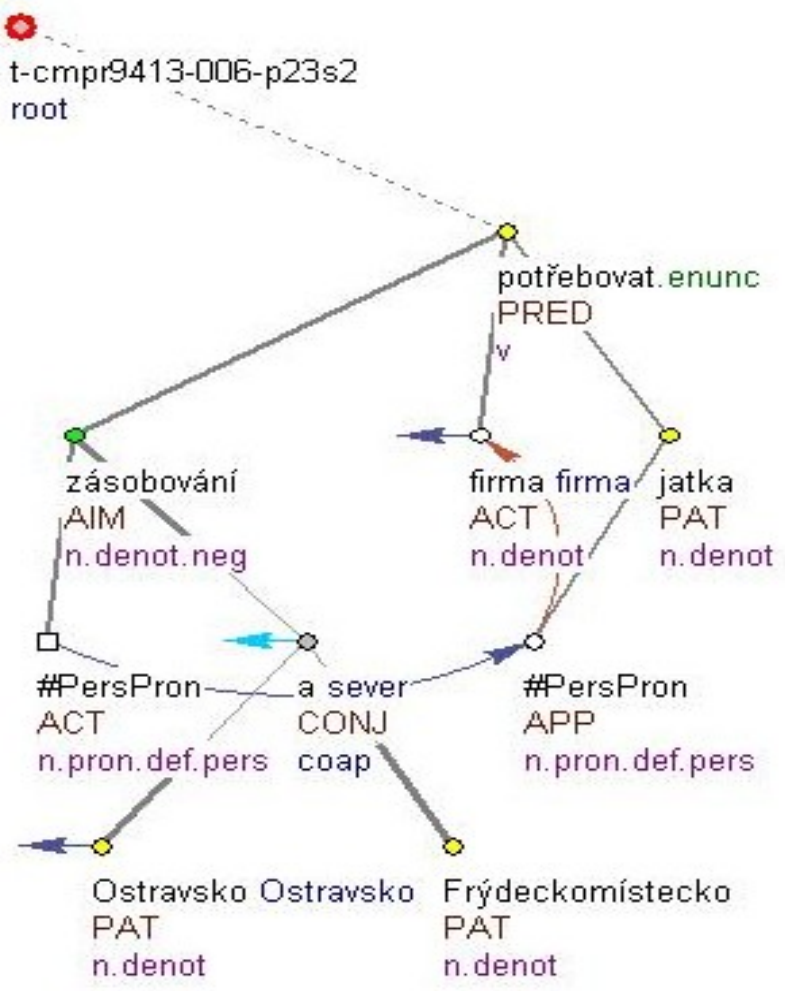
DĚLÁ SE

Technické provedení v PDT

odkaz na ID antecedenta v atributech:

- **coref_gram.rf –** pro zaznamenání gramatické koreference
- **coref_text.rf (+ informal-type: 0, NR) –** pro zaznamenání zájmenné a jmenné textové koreference
- **coref_special (segm, exoph) –** pro zaznamenání případů speciálních druhů odkazů: na segment předcházejícího textu (segm) a exoforický odkaz mimo text (exoph)
- **bridging (+ informal-type: SET, PART, FUNCT, CONTRAST, REST) –** pro zaznamenání asociační anafory

Reprezentace v TrEdu



Pro zásobování Ostravska a Frýdeckomístecka potřebuje firma svá jatka .

Teoretické zásady anotace koreference

- princip řetězce (odkaz na nejbližší předcházející koref. uzel)
- princip maximální délky koref. řetězce (propojení gramatická k. — textová zájmenná k. — textová jmenná k.)
- maximální velikost koref. členů (vždy celý podstrom)
- princip kooperace s anotací TG-roviny (neanotujeme apozici, predikaci, přímé potomky s funktory APP, PAT, MAT a AUTH atd.)
- … princip rozhodujícího koreferenčního vztahu (před anaforickým), princip zvláštní váhy podílu na kohezi textu, princip preference koreference před asociační anaforou a jiné

Gramatická koreference

- koreference zvratných zájmen, hlavně *sebe, svůj* *(Sobě nedopřeje matka nikdy nic.)*
- koreference vztažných prostředků *který, jenž, kam,kde* apod. *(Člověk, který chce všechno.)*
- koreference v recipročních konstrukcích *(Sultáni se vystřídali {#Rcp.PAT} na trůnu.)*
- kontrola *(Začít {#Cor.ACT} číst.)* a kvazikontrola (*Karel podal {#QCor.ACT} stížnost policii.)*
- koreference u doplnění s dvojí závislostí vyjádřených slovesnou formou

Textová koreference zájmenná

Identita referentů. Koreferující člen je:

- osobní nebo přivlastňovací zájmeno v 3. osobě *(Helena poprosila maminku, aby na ni počkala)*
- ukazovací zájmeno TEN v substantivní funkci (*Helena poprosila maminku, aby na ni počkala. Ta to ale odmítla.*)
- při aktuální elipse, kdy je do tektogramatického stromu doplněn nový uzel se zástupným t-lematem #PersPron (*Helena poprosila maminku, aby* #PersPron *na ni počkala.*)

Antecedent nemusí být ve stejné větě s koreferujícím členem. Antecedent se určuje na základě kontextu

Anotace gramatické a zájmenné textové koreference

ručně a částečně automaticky na celém PDT

- gramatická + textová:
  - r. 2002 — 2004
  - 2 anotátorky
  - **46 242** linků v 49 431 větách (50.3% gram.k., 48.4% text k.)
  - automatická předanotace

Textová koreference jmenná

- *Helena poprosila **maminku**, aby na ni počkala. **Maminka** však řekla, že nemůže.*
- ***Helena** poprosila maminku, aby na ni počkala. Maminka však prosbu **dcery** ignorovala.*

  Ale

- *Dcera by měla poslouchat maminku. Například moje dcera mě poslouchá.*

  nejsou koreferenční

Textová koreference jmenná slovní druhy

Koreferující člen může být:

- **substantivum** (*Helena*/*ona*/*#PersPron, ... — Helena*)
- **příslovce** *(v Praze — tam)*
- **adjektivum** *(*přivlastňovací typu *otcův* nebo pokud je odvozeno od NE - *německý)*

  + v některých případech také číslovka, sloveso

Dodržování koreferenčního řetězce

Helena poprosila maminku, aby #PersPron na ni počkala. Maminka však řekla, že #PersPron nemůže.

- řetězec se dodělává automaticky: A ← B ← C
- problémy:

"...očekávají návštěvu sparťanského prezidenta Macha s manažerem Nehodou, kteří by měli podat vysvětlení. (...) Musí zasáhnout manažer nebo prezident klubu."

… manažer = prezident

(naštěstí jediný případ a dá se pak vyhledat automaticky)

Typologie TKR vztahů

rozlišujeme specifickou a nespecifickou referenci

- **TYP 0** (specifická reference)
  - *SYN (různé nominace, např. *Helena - slečna*)
  - *ER (hyperonym antecedenta, např. *mango — to ovoce*)

např.: *Helena — #PersPron — slečna — s ní — dcera*

- **TYP NR** (nespecifická reference)

např. *Tímto faktorem je podnikatel - inovátor , který se snaží o zisk , a proto logicky nemůže existovat ve stavu statiky , která nezná ani zisk , ani ztrátu. Tento podnikatel se od manažera liší tím , že …*

! sem patří většina případů koreference abstrakt, pokud není nápadně jinak

Bridging anaphora

Helena vstoupila do místnosti. Ze stropu kapala voda.

- nejde o koreferenci, ale o sémantický vztah
- podílí na kohezi textu
- neudržuje se řetězec s jinými typy koreference
- snažíme se propojovat jenom autosémantické uzly

Typologie bridging vztahů

- **část — celek** (významy PART_WHOLE a WHOLE_PART v atributu bridging v informal-type)
- **množina — podmnožina/element množiny**(SET_SUB a SUB_SET)
- **funkce — objekt** (P_FUNCT a FUNCT_P)
- vztah sémantického **kontrastu** (CONTRAST)
- **ostatní** (REST)

Bridging - PART

část — celek (PART_WHOLE a WHOLE_PART)

- *Dělal jsem bez přestávky celé **týdny** , často v noci.*
- *Německo – Bavorsko – Mnichov*

Neanotujeme, pokud ACMP, PAT, APP, MAT, AUTH:

- *Na břehu Starnberského jezera u místa utonutí byla postavena **kaplička** s královými daty narození , vlády a smrti , s křížem a mramorovou pamětní deskou.*
- ***strop** této místnosti*

Bridging - SET

množina — podmnožina / element množiny (SET_SUB a SUB_SET)

*Na rozdíl od dobře vybaveného FS dnes nikdo z téměř dvou stovek **poslanců** kromě předsedy a místopředsedů sněmovny a šéfů jejích výborů nemá svou kancelář , pracovní stůl , židli a telefon .*

také u jmenných skupin s nespecifickou referencí:

Nový VW Golf je vybaven motorem o síle..." - "Dostali jsme možnost se novým golfem projet.

Hraniční případy SET a PART

občas rozlišení ovlivňuje pouze počitatelnost jména

*Jeho hlavní výhodou by mělo být lepší napojení na televizní přenosovou **techniku** : zatímco dnes přenosové vozy {SET/PART} blokují parkovací prostor před starou sněmovnou , v budoucnu zajedou do Thunovské a kabely {SET?} se snadno spojí s tiskovým centrem .*

*Když ho smrt překvapila u psacího stolu, revidoval právě **text** Prezidentské adresy, kterou pronesl několik dní před tím v Americké ekonomické asociaci . Poslední věta { PART?}, kterou v životě napsal , zněla : Stagnacionisté se mýlí v diagnóze důvodu , proč by kapitalistický proces měl stagnovat .*

Bridging - FUNCT

funkce — objekt (P_FUNCT a FUNCT_P)

- *Někteří se vrátili do Německa, další přešli na **faru** v nedalekém Jeníkově a spojili se se skupinou , která tady byla již ubytována. Podobné zkušenosti jako v Oseku potvrdil i farář v Jeníkově pan Matfiak .*

- *ministr — vláda* (SET) vs. *premiér — vláda* (FUNCT)

Bridging — CONTRAST

vztah sémantického **kontrastu** (CONTRAST)

- *A přesvědčen jsem ještě o jednom - je třeba mít vysoké **cíle** a s malými [cíly] se nespokojit .*
- ***Lidi** nežvýkají , to jenom krávy.*

Neanotujeme, pokud jsou propojeny ADVS:

- *Dočasný podnikatelův **zisk** bude anulován , ale trvalý zisk z jeho inovace zůstane zachován společnosti ve formě nižších cen nebo technicky dokonalejších výrobků .*

NB! Zajímavé prolínání s AČV a diskurzivními vztahy.

bridging - REST

ostatní přesněji nezařazené případy (REST)

- vztah rodinné příslušnosti (*děda — vnuk*)
- místo — obyvatel (*Mexiko — Mexičán*)
- autor — dílo (pokud není AUTH) *(obraz — autor)*
- stejné nominace s podílem na kohezi (*přihrála náhoda - do hry vstoupila další náhoda*)
- událost — argument (*podnikání — podnikatel*)
- některé jiné ale bez fanatismu

anafora bez koreference

- *"Duha?" Kněz přiloží prst k tomu slovu, aby nezapomněl, kde skončil.*
- *Protože tenhle Adolf Hitler nebyl vůdce velkoněmecké říše, ale pták druhu tučňák královský. A to jméno dostal vlastně dodatečně.*
- *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.*
- *Jak se Vám zamlouvala Pragobanka Cup? - Takovou/podobnou/stejnou akci bychom také uvítali*

Zatím patří do REST. Přemýšlíme o zavedení Bridging-ANOF

Speciální typy odkazů **nové** exoph

- *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.*

- *A tu se dostáváme zpět k počátku tohoto textu .*

Speciální typy odkazů nové segm

- *Celní unie bude sice existovat na papíře ještě dalších dvanáct měsíců, ale v praxi dostanou vzájemné vztahy punc tvrdosti mezinárodního obchodu . Poroste administrativa. Jistotu v tomto směru dávají nejnovější kroky vlády SR , která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience .*
- *Tak jako každý Mexičan, i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle, a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku . Právě v té době přihrála náhoda Santa Annovi do cesty Thomase Adamse , fotografa a především vynálezce všeho druhu .*

nový `segm` - 2

typ `segm` dáváme, pokud struktura stromu nedovoluje odkázat na náležitý podstrom:

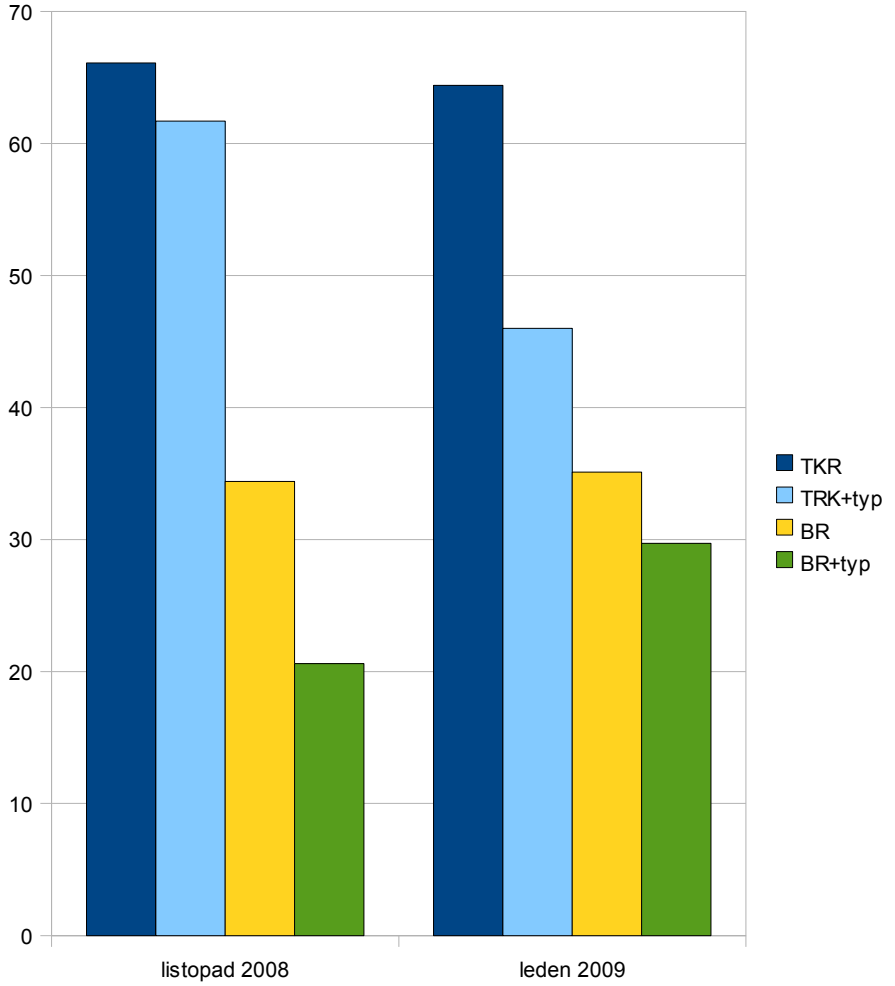
- *Od 1 . dubna nebude ÚNMS SR rozhodnutí české zkušebny potvrzovat .Tato funkce přejde na příslušnou slovenskou zkušebnu.*
- *Chtěl jsem být největším ekonomem na světě , největším milencem na světě a největším jezdcem na světě .Vzhledem k pokročilému věku třetí cíl už nestihnu .*

Technické provedení

- standardní rozšíření TrEdu (bridging extension)
- 20 vět před aktuálně anotovanou větou a po ni (dá se nastavit jinak)
- zvýrazňují se stejná lemmata (podtržením)
- zvýrazňují se (barevně) všechny vztahy s aktuálním uzlem (podle typu vztahu - modře, bledě modře a červeně)
- anotovat se může na stromu nebo na textu
- problém u velkých souborů (nad 30 vět)
- předanotace NE (německý - Německo)

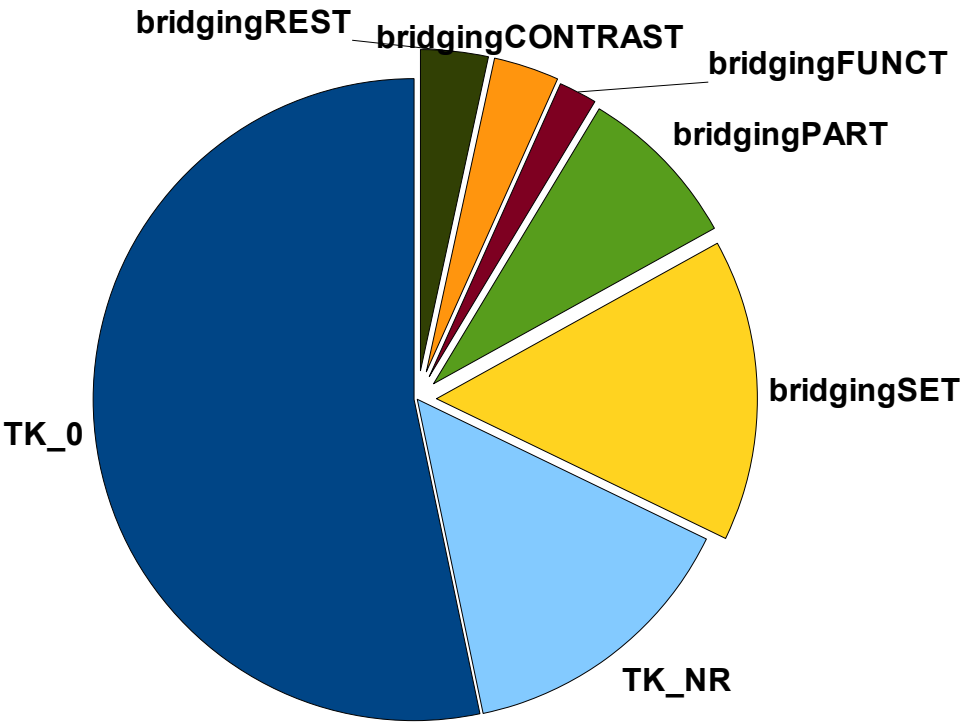
průběh anotace
mezianotátorská shoda

- začátek 10.2008
- 10.-12.2008: 3 anotátoři
- 1.2009 - ...: 2 anotátoři
- 1000 vět měsíčne
- pokud stejně - ½ PDT ke konci roku



statistika dat

- 492 souboru z train-1 a train-2: 7434 vět (15% PDT), 124 648 slov
- 3493(gram) +3282(text)=6775 starých šipek
- 8597(text) +4029(bridging)=12626 nových šipek
- 3282(pron.text) vs. 8597(noun.text)



problematická místa: **předložkové fráze**

struktura TGS neumožňuje rozlišovat NP a PP

- *v době karnevalu — v těchto dnech — karneval*

problém: den KOREF karneval

Německá stávka v době karnevalu... Že mnohé Němce bude v těchto dnech bolet hlava, s tím se počítalo. Ve čtvrtek přece začal masopustní karneval...

- *za Prahou - části města — tu, před válkou — po válce*

nebo naopak koreferenční jsou pouze PP:

- *před začátkem utkání — při rozcvičování*

problematická místa: **koreference na kontejner**

*Tří a půl **tisíce** dělníků vyhlásili stávku. Stávkující žádají zvýšení platů o šest procent. Do 8. března se počet stávkujících může zdvojnásobit.*

nebo

*Tří a půl **tisíce** dělníků vyhlásili stávku. Stávkující žádají zvýšení platů o šest procent. Do 8 . března se počet stávkujících může zdvojnásobit .*

problematická místa: **abstraktní jména**

Míra nezaměstnanosti by se měla vyvíjet protikladně , než ve standardní ekonomice . […] růst nezaměstnanosti v letech 1991 — 1993 značně zaostal za poklesem HDP. Pokračující privatizace a restrukturalizace si však vynutí zvýšení míry nezaměstnanosti z 3.5 % koncem roku 1993 na 5 — 6 % ke konci příštího roku.

- anotovat — neanotovat? těžko stanovit hranice koreference
- pokud ano, jako typ 0 nebo typ NR? (teď jako NR, pokud tomu není nápadně jinak)
- problém hranice mezi konkrétními a abstraktními NP (*zisk*)
- podobně u dějových jmen

PLÁNY

- pokračovat v anotaci
- hledat a vyučovat nové anotátory
- vylepšovat mezianotátorskou shodu
- srovnávat a vyhodnocovat materiál
- základy pro automatickou předanotaci
- začít strojové zaučování, creating features