Anotace textových vztahů v Pražském závislostním korpusu

Zuzanna Bedřichová, Lucie Mladová
ÚFAL MFF UK

Seminář formální lingvistiky 9. 3. 2009

Na úvod

- základní východiska (rekapitulace)
- teoretická a technická řešení anotace na podkladě PDTB 2.0
- návrhy na základě první pokusné anotace

Na úvod

- základní východiska (rekapitulace)

  text – promluva – diskurz

  propozice – *abstract object* – elementární predikační struktura

Na úvod

- „textové“ vztahy = prostředky textové koherence:
  - koreference a bridging
  - AČV, tematické posloupnosti
  - grafické či zvukové členění
  - syntaktická struktura textu
  - rétorické (kompoziční) vztahy
  - komunikační a pragmatické faktory
  - …

Na úvod

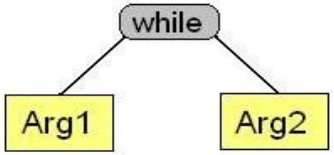
- „textové“ vztahy = prostředky textové koherence:
  - koreference a bridging
  - AČV, tematické posloupnosti
  - grafické či zvukové členění
  - **syntaktická struktura textu**
  - **rétorické (kompoziční) vztahy**
  - komunikační a pragmatické faktory
  - …

Na úvod

- „textové“ vztahy = prostředky textové koherence:
  - koreference a bridging
  - AČV, tematické posloupnosti
  - grafické či zvukové členění
  - **syntaktická struktura textu**
  - **rétorické (kompoziční) vztahy**
  - komunikační a pragmatické faktory
  - …
- úvodní přednáška - 26.11. 2007 Šárka Zikánová
  - Tektogramatická reprezentace v PDT 2.0
  - Penn Discourse TreeBank 2.0

Penn Discourse TreeBank 2.0

- Institute for Research in Cognitive Science, University of Pennsylvania
- Aravind Joshi, Rashmi Prasad, Alan Lee, Eleni Miltsakaki, Bonnie Weber a další
- verze 2.0 – únor 2008 v LDC, cca 49 000 vět z WSJ
- princip anotace:
  - textový konektor (*discourse connective*) jako predikát binárního vztahu
  - argumenty – propozice (*abstract objects* dle Ashera 1993)
- *John eats porridge for breakfast*, while **Mary eats muesli.**



Penn Discourse TreeBank 2.0

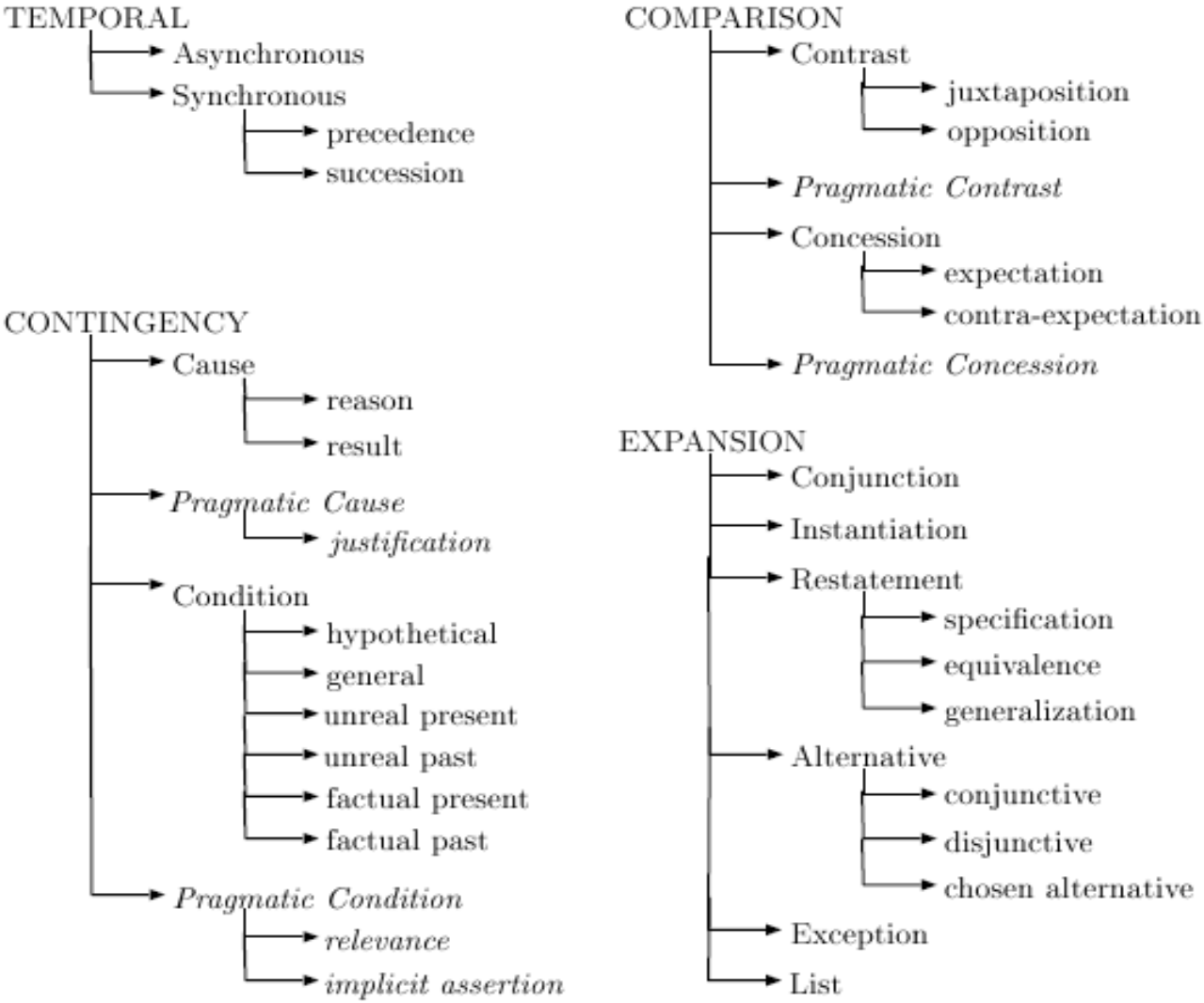
- verze 2.0:
  - anotace všech **explicitních** (povrchově přítomných) konektorů v PTB, anotace jejich argumentové struktury (**rozsahu argumentů**)
  - anotace **typu významového vztahu** mezi argumenty explicitně vyjádřených vztahů
  - anotace **implicitních** konektorů (vkládání vhodného konektoru)
  - anotace **typu významového vztahu u implicitních** vztahů – provedeno mezi všemi sousedními větami kromě hranic odstavců

Penn Discourse TreeBank 2.0

- doplňující anotace:
  - **AltLex** – alternatively lexicalized relation *„the reason is…“, „v každém případě…“*
  - **EntRel** – entity based relation - koreference
  - **NoRel** – tam, kde nebyli schopni zachytit žádný vztah

  - **attribution** – přiřazení obsahu propozice původci - pisateli/mluvčímu nebo někomu jinému *„according to“, he admitted“*
    - Není skutečným „discourse“ vztahem, netýká se vztahu konektor – argument, přesto se podílí na výstavbě textu

Druhy významových vztahů v PDTB 2.0



Pro zajímavost: Hindi Discourse Relation Bank

- A. Joshi, Rashmi Prasad, Samar Hussein, D. M. Sharma (Penn a LTRC Hyderabad India)
- vznikající projekt, navazuje na závislostní anotaci, 1 milion slov v Hindi – další projekty: čínština, turečtina
- úprava anotačního scénáře PDTB 2.0
  - úprava souboru významových značek
  - samostatný atribut „prag“ pro pragmatické významy
  - implicitní vztahy anotovány nejen mezi sousedními větami, ale kdekoli, kde mají smysl
  - jiná práce s konektory - možná je např. inference AltLexu – tj. alternativního vyjádření konektoru
  - neanotují podřadicí spojky (tato informace obsažena na rovině syntaxe)

K čemu „diskurzní“ korpusy?

- automatická sumarizace textu
- získávání a odvozování informací
- dialogové systémy, vztah otázka – odpověď
- automatická anotace dalších korpusů
- podklad pro lingvistické analýzy založené na korpusech (první korpus textových vztahů zaměřený na češtinu)
- ?? strojový překlad

Komputační experimenty s PDTB a ostatními diskurzními korpusy (RST Treebank, Discourse Graphbank)

- Emily Pitler a Annie Louis Nenkova
- experimenty s distribucí typů vztahů a konektorů, s automatickou sumarizací, extrakce disk. anotace pro rysů na „predicting coherence“ (30 % koherence zajišťuje diskurz), dále snahy o automatizaci anotace
- jako užitečné se ukázaly dvě věci:
  - mít diskurzní anotaci propojenou s koreferencí
  - „whole structure is more helpful“
  - anotovat typ textu
- k sumarizaci byly relevantní explicitní vztahy (konektory), byly daleko aktivovanější než implicitní
- běží další experimenty: s atribucí, information extraction atd.

Osnova

- základní východiska (rekapitulace)
- teoretická a technická řešení anotace na podkladě PDTB 2.0
- návrhy na základě první pokusné anotace

Textová anotace v Praze

- Eva Hajičová, Šárka Zikánová, Zuzanna Bedřichová, Lucie Mladová, Jiří Mírovský, Zdeněk Žabokrtský, Pavel Češka

- součást *GAČRu 2009 - 2011*

  Od struktury věty k textovým vztahům – především pro korerenci a aktivovanost (Bára Hladká, Anja Nedoluzhko)

- demo anotace pro PDT 2.5, malý vzorek dat

Textová anotace v Praze

- Co poskytuje tektogramatika
  - v rámci stromu
    - závislostní hrana určitého typu (COND,CNCS, CAUS, …)
    - koordinace
    - PREC
  - i mezi stromy
    - anotace rozšířené koreference (Anja Nedoluzhko)
    - kopírování uzlů z předchozích vět u elipsy

Lingvistické předpoklady pro anotaci

- revize pojetí hypotaxe a parataxe
- syntax věty není stejné povahy jako významová výstavba textu
- některé závislostní hrany a koordinace s jejich sémantikou na TR můžeme v zásadě převzít, důraz na vztahy **mezi** stromy

Parataxe a hypotaxe

- pro klasifikaci významových vztahů mezi propozicemi v textu – (další) odhlédnutí od formálního vyjádření
- příklady
  - přípustka – adverzativní vztah
  - příčina – důsledek
  - předčasnost – následnost

- podmínka jako typ textového vztahu:

  Usmažím palačinky. Musíš mi ale nejdřív koupit vajíčka.
- výjimka jako typ textového vztahu:

  Nevěnuji se žádnému sportu. Jen si občas chodím zaplavat.

Větná syntax a výstavba textu

- syntax věty není stejné povahy jako významová výstavba textu
  - V rámci textu lze najít mnoho vztahů v tradičním pojetí nesyntaktických (restatement - v PDTB některé ze třídy *Expansion*)

Nikdy netrávila večery doma. Chodila například na procházky s přáteli. exemplifikace

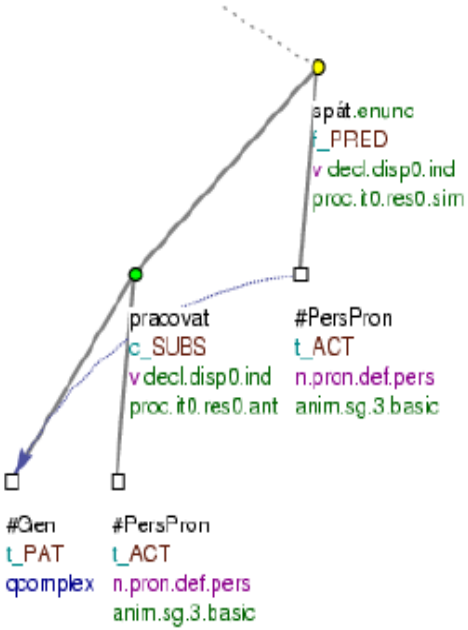
Jel do zatáček opatrně. Vždy si nadjížděl. specifikace

Metoda záleží jenom na vás. Prostě to udělejte podle sebe. ekvivalence

Větná syntax a výstavba textu II.

- Některé závislostní vztahy se nepodílejí na výstavbě roviny textové
  - DIFF
    - *Čím je víno starší, tím je lepší.*
- Některé mají deiktický prvek, který zachycuje anotace koreference
  - MANN – pouze s odkazovacím slovem
    - *Opravil ledničku kladivem.*
    - *Opravil ledničku. Udělal **to** kladivem.*

    specifikace



  - SUBS
    - *Místo toho, aby pracoval, spí.*
    - *Měl by pracovat.* (Nepracuje.) ***Místo toho** spí.* rektifikace

Mezivětné (mezistromové) vztahy

- závislostní hrany a koordinace s jejich sémantikou na TR můžeme v zásadě převzít, důraz na vztahy **mezi** stromy
  - Čili: *John eats porridge for breakfast*, while **Mary eats muesli.**
    - již zachyceno v rámci TR

- zároveň je třeba rozšířit některé zásady TR pro potřeby anotace textu:
  - konjuktivní alternativa, větné apozice, nepravé věty vedlejší atd.

- nechceme přepisovat TR, jen přidávat nové informace

Soubor textových významových vztahů pro ČJ

TEMPORAL	CONTINGENCY	COMPARISON	EXPANSION
synchronous	**cause** (reason + result)	**confrontation** (PDTB juxtaposition)	**conjunction**
asynchronous	**condition**	**opposition**	**instantiation**
	purpose	**restrictive opposition** (+ exception)	**specification**
		concession	**equivalence**
		replacement = correction + substitution (PDTB chosen alternative)	**generalization**
		gradation	**conjunctive alternative**
			disjunctive alternative
			list

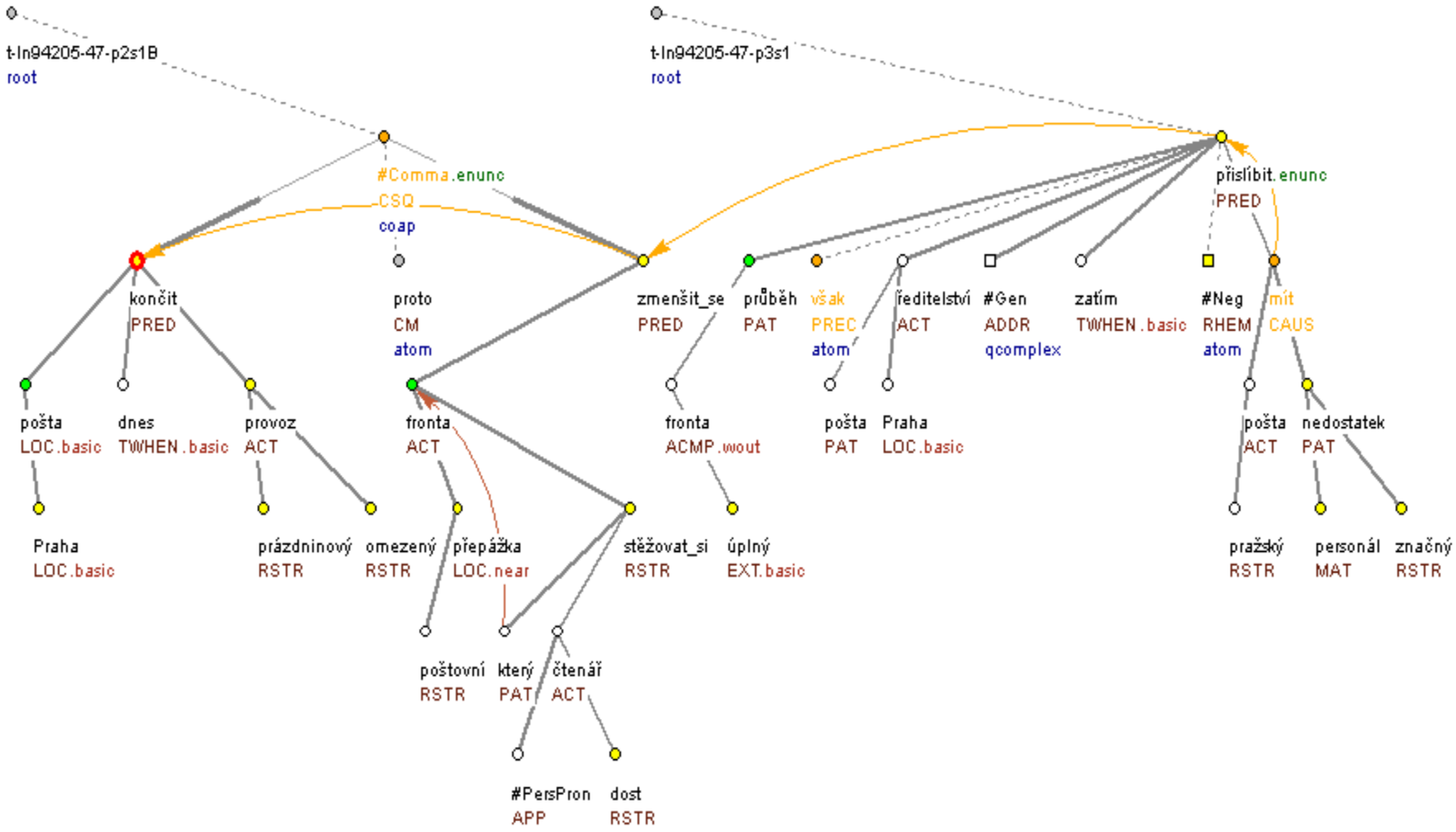
Porovnání anotačních principů v PDTB a PDT

	PDTB	PDT
rovina	discourse level	„vrstva“ textových vztahů není samostatná rovina vyšší k TR
konektor	discourse-level predicate	abstrahování od formy: konektor je jedno z možných vyjádření textového vztahu
anotovány vztahy	explicitní a implicitní	(zatím?) jen explicitní
atribuce	anotována	zatím nebude anotována
pragmatické významy	anotovány	zatím nebudou anotovány
argumenty	2	2 i více
anotace	na textu	na textu, ale s promítnutím na stromy
vztahové dvojice typu „příčina“ – „důsledek“	anotovány zvlášť, tj. dvěma vztahy	sloučeny
možnost označení dvěma významy	ano	ano

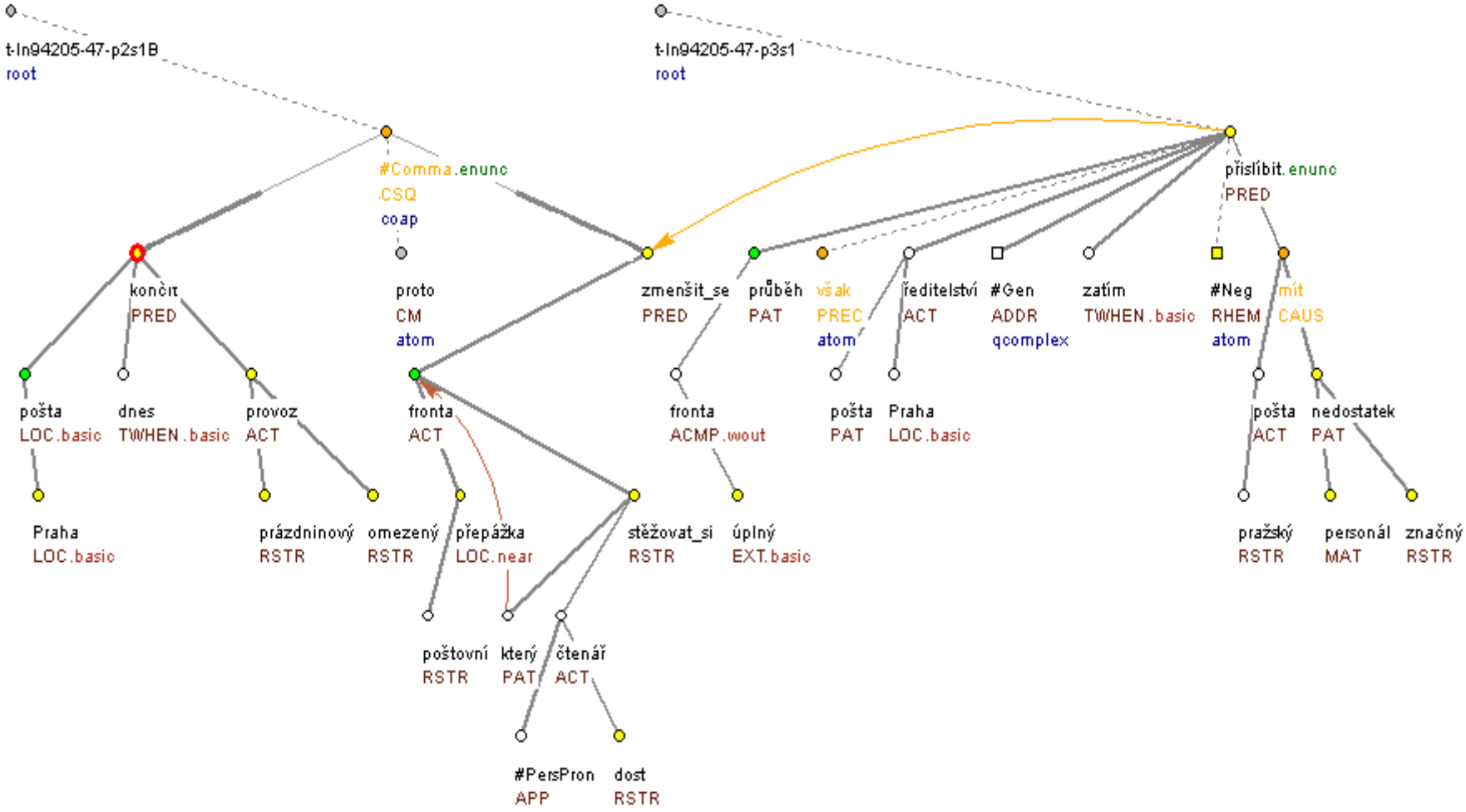
Technická stránka anotace

- anotace v TrEdu v novém kontextu discourse
  - textové okno + okno se stromy
  - podobně jako koreference: mezi kořeny podstromů, které odpovídají dvěma argumentům jednoho vztahu, se kreslí „diskurzní“ oranžová šipka
  - tektogramatika není porušována
  - technický uzel nebo jiný prostředek zachycení vnořených a seznamových vztahů (např. Arg1 = 3 stromy)
  - viditelné hranice odstavců původního textu
  - směr šipky vypovídá o orientaci vztahu
    - asymetrické: příčina ← důsledek
    - symetrické: např. konfrontace, šipka vede konvenčně k Arg1 (první v textu)

Technická stránka anotace



Technická stránka anotace



Technická stránka anotace

The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said.
Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

A Lorillard spokewoman said, "This is an old story.
We're talking about years ago before anyone heard of asbestos having any questionable properties.
There is no asbestos in our products now."

Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes. "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute.

Conn/AltLex | Conn/AltLex Attr | Arg1 | Arg1 Attr | Arg2 | Arg2 Attr | Sup1 | Sup2

Anotace v textovém okně s projekcí do stromů

- výhody anotace v textovém okně (s projekcí do stromů)
  - výrazně jednodušší značení rozsahu argumentů
    - vsuvky, vícevětné přímé řeči atd.
  - možnost označení konektoru
  - neovlivňování stromy (zejména pro rozsah argumentů)

Anotace v textovém okně s projekcí do stromů

- výhody okamžité projekce do stromů
  - kreslí se okamžitě struktura celého dokumentu
  - možnost zobrazit koreferenci i námi anotované textové vztahy najednou
  - možnost využití technického uzlu (seznam, vnořené vztahy)

- nevýhody
  - T-stromy mohou ovlivňovat anotátora při výběru rozsahu argumentu (ale mohou jej i korigovat!)
  - rychlost TrEdu?

Osnova

- základní východiska (rekapitulace)
- teoretická a technická řešení anotace na podkladě PDTB 2.0
- návrhy na základě první pokusné anotace

První pokusná anotace

- 3 anotátorky, 1 text (34 vět), explicitně i implicitně vyjádřené vztahy, kreslení šipek na t-stromech + přiřazení vztahu
- Shoda: označeno 31, 30, 21 vztahů
  - šipka mezi 2 stejnými uzly: 2 – 8
  - významové značky: max 3
- pro srovnání - shoda v PDTB – na 2 úrovních:
  - Class level 92%
  - Type level 77%
- největší problémy
  - kde vztah je a kde není při absenci explicitního konektoru?
  - rozsah argumentů – při anotaci pouze na stromech
  - velké množství inferencí
  - orientovanost šipek

První pokusná anotace

Poučení:

- naprostá nutnost vycházet z konektorů
  - zatím ne implicitní vztahy
  - mít soupis konektorů a neodchylovat se od něj
- možnost přiřadit 2 značky
- anotovat primárně na textu

Ideální postup anotace

- anotátor anotuje na textu:
  - argumenty a konektor
- do stromu se promítne:
  - argument jako množina uzlů
  - mezi množinami (nikoli mezi dvěma uzly) se vytvoří šipka a rovnou se „zeptá“ na vztah
- každý anotátor anotuje vždy celý text, ne pouze vybrané konektory (jako v PDTB)
- začít od anotace explicitních vztahů (průběžně vytvářet soupis konektorů a AltLexů)
- hierarchická soustava vztahů – dvě úrovně textových značek (pro měření shody na hrubší a jemnější úrovni)

Rovina (pra)lesa