

Corpus Annotation – Sentence and Discourse

3. Coreference in the Sentence and in the Text

PDT layers and coreference

- The three PDT layers – capture grammatical information
- Coreference relations – textual relations – “beyond” grammar

BUT:

the aim: by annotating these relations to get more insight into the inter- and intrasentential structure

Annotation of coreference relations in PDT

- coreference relations in the narrower sense
- a binary relation between an anaphor and an antecedent:
 - the antecedent may be in a different TGTS
 - the antecedent may also be an entity that is not represented in any TGTS
- 2 kinds of coreference
 - grammatical
 - textual

Principles of coreference annotation (1)

- Chains – reference to the nearest antecedent
- Maximal length of chains (incl. grammatical and textual coreference)
 - Example:
 - Anička poprosila svou maminku, aby na ni počkala. Matka řekla, že jde do divadla.
 - Anne asked her **mother #PersPron** to wait for her. **Mother** said that **#PersPron** goes to the theatre.
 - the chain is established automatically A <- B <- C <- D

Principles of coreference annotation (2)

- Maximal “scope” of the units: whole subtree
- “cooperation” with the TGTS’s: no special annotation of apposition, predicates etc.
- Additional considerations: coreferential relation preferred to anaphoric and to associative anaphora, contribution to the coherence of the text ...

Grammatical coreference

- verbs (nouns, adjectives) of control or quasi-control
 - *John asked Mary to [0] come.*
 - *John submitted a [0] complaint to the police.*
- reflexive pronouns
 - *John shaved **himself**.*
- relative pronouns
 - *John, **who** came late, apologized.*
- verbal complements
 - *John came [0] bare-footed.*
- reciprocity
 - *John and Mary kissed [0].*

Textual coreference

- Present stage:
 - in the whole PDT 2.0
 - demonstrative and anaphoric pronouns (also in their zero form), 3rd person
 - bridging anaphora is not included
 - in a sample of 80 PDT documents
 - anaphoric relations leading from nouns incl. a rough classification of bridging anaphora

Types of textual coreference

- link to a particular node
- link to the governing node of a subtree
- segm(ent): referent is a whole segment of text
- exoph(or): referent is „out“ of co-text
- unsp(ecified): reference is difficult to be specified

Link to a particular node

- this node represents an antecedent of the anaphor:

*Do you think that the decision of NATO whether **[it]** will be enlarged or not will depend on the attitude of Russia?*

→ the link from ***it*** leads to **NATO**

Link to the governing node of a subtree

- antecedent is represented by this node plus (some of) its dependents; also the way how a link to a previous/following clause or a whole previous sentence is being established:

*But it is a different thing when someone is an entrepreneur and then goes into politics than when political changes elevate somebody to the top and he then uses **this** in his economic activities.*

→ the link from **this** points to the root of the tree (*elevate*) = to the main verb of the second conjunct.

Segm(ent)

- referent is a whole segment of (previous) text larger than one sentence (phrase):

*According to Kohl it should not be forgotten that on June 22, 1941 Germany attacked the Soviet Union. Germans on behalf of Germany caused the Russians to suffer immensely. It also cannot be forgotten what the Russians did to Germans. From all **this** we should learn.*

Segm(ent) 2

- includes also the cases, when the antecedent is understood by inferencing from a broader co-text:

*The big shots buy in a bank for ten and sell for fifteen. But this leads to a rapid transformation. The acrages of about 25 ha disappear, the number of owners raises to 500. I guess that within two years they will be able to pay back the debt to the bank and in the third year they will work for themselves. And they will hire only capable people, it will be in their best interest. Those who understand **this**, will have an advantage.*

Exoph(or)

- a specifically marked link denoting that the referent is “out“ of the co-text, it is known only from the situation:

*In the height of summer 1939 only a few people could believe the hopeful words Chamberlain uttered [...] after the return from Munich: I think that **this** is peace for our time.*

→ **this** = Munich Treaty

Unsp(ecified)


- a specific mark reserved for cases of reference difficult to be identified; a decision is not to be made between two or more referents but that the reference cannot be specified even if the situation is taken into account:

*The disappearance of the medical instrument weighing 700 kg **[they]** announced on June 30th this year. According to the information of LN, however, the radiator disappeared by the end of the last year.*

Typology of relations

- TYPE 0: specific reference
 - synonymy: *Helen – the girl*
 - hyperonym: *orange – the fruit*
- TYPE NR: non-specific reference
 - *This factor can be illustrated by an **entrepreneur-manager**, who wants a profit, and therefore he logically cannot exist in a static state, which doesn't know any profit or loss. Such **an entrepreneur** differs from a manager who ...*

Bridging anaphora

- *Helena vstoupila do místnosti. Ze stropu kapala voda.*


The diagram consists of a rectangular box with the text 'WHOLE_PART' inside. An arrow points from the top-left corner of the box up to the word 'místnosti' in the text above. Another arrow points from the top-right corner of the box up to the word 'stropu' in the text above.
- Helen entered **the room**. From **the ceiling**, water was dropping ...
- No coreference, but a semantic relation, contributing to the coherence of the text
- no chain with other types of coreference is established

Types of bridging relations

- **Part – whole** (PART_WHOLE and WHOLE_PART in the attribute ‘bridging’)
- **set — subset/element of set** (SET_SUB and SUB_SET)
- **Function - object** (P_FUNCT and FUNCT_P)
- **semantic contrast** (CONTRAST)
- **other** (REST)

Bridging - PART

- *Dělal jsem bez přestávky celé týdny , často v *nocí*.*
I have been working without a break whole weeks, often at night.
 - *Německo – Bavorsko – Mnichov*
Germany – Bavaria - Munich
- Not annotated: ACMP, PAT, APP, MAT, AUTH:
- *strop této místnosti*
the ceiling of this room

Bridging - 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 .*
 - In contrast to ... none of the 200 hundred of **MP's** except for the **chairman** and **vice-chairmen** ... and **the chairs** of the committees ...
- also with noun groups with a non-specific reference:
- *Nový VW Golf je vybaven motorem o síle..." - "Dostali jsme možnost se novým golfem projet.*
 - New **VW Golf** is equipped ... We have got the chance to drive the new **golf**.

Bridging - FUNCT

- a difference is made:
- *ministr* — *vláda* (SET)
 - minister - government
- *vs. premiér* — *vláda* (FUNCT)
 - prime-minister - government

Bridging — CONTRAST

- *A přesvědčen jsem ještě o jednom - je třeba mít vysoké cíle a s malými [cíli] se nespokojit .*

And I am sure about one thing: it is necessary to have lofty aims and not to be satisfied with small (ones).

Not annotated if the dependency relation is ADVS:

- *Dočasný podnikatelův zisk bude anulován , ale trvalý zisk z jeho inovace zůstane zachován společností ve formě nižších cen nebo technicky dokonalejších výrobků .*
 - *The transitory ... profit will be annulated but the permanent profit ...will be preserved ...*

Bridging - REST

Other relations not specifically distinguished (REST)

- Family: *grandfather - grandchild*
- Place – inhabitant: *Mexico - Mexican*
- author — work: *picture - author*
- same denomination to support cohesion of the text: *a chance helped – another chance entered the game ...*
- event — participant of the event: *enterprise - entrepreneur*

Anaphora without coreference

- "Duha?" Kněz přiloží prst k tomu slovu, aby nezapomněl, kde skončil.
 - "A *rainbow?*" The priest pointed to the *word* ...
- Protože tenhle Adolf Hitler nebyl vůdce velkoněmecké říše, ale pták druhu tučňák královský. A to jméno dostal vlastně dodatečně.
 - Because this *Adolf Hitler* was not the leader of ... but a bird ...
And he got the *name* additionally, ...
- Jak se Vám zamlouvala Pragobanka Cup? - Takovou/podobnou/stejnou akci bychom také uvítali
 - How did you like the *Pragobanka Cup?* We would welcome a similar *event* ...

Currently: REST

Annotational scheme

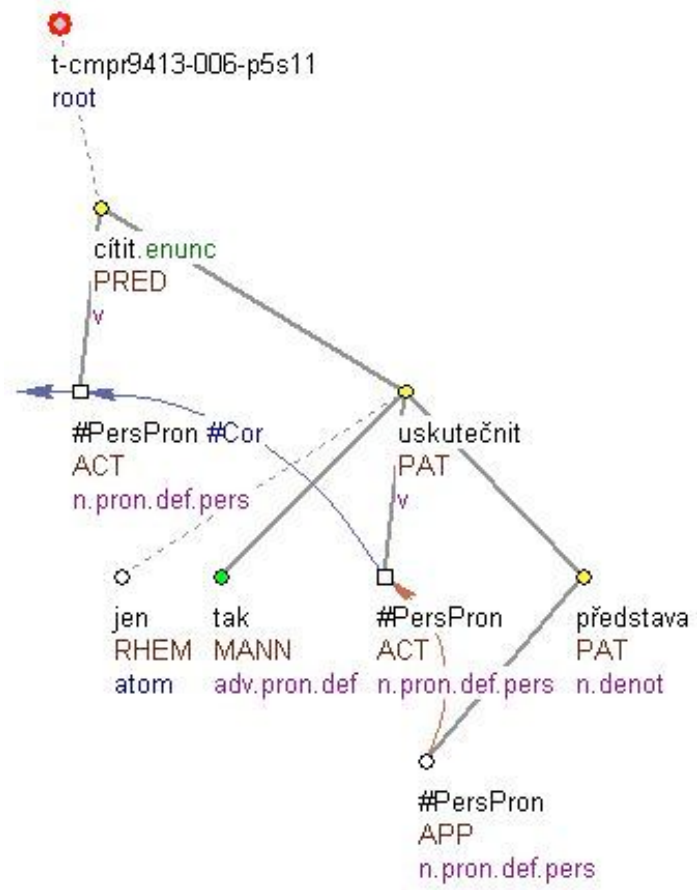
- explicit coreference links are technically represented as pointers (reference) leading from anaphor t-nodes to their antecedent t-nodes
- three coreferential attributes with an anaphor:
 - **coref_gram.rf** – identifier (or a list of identifiers) of the antecedent(s) in the sense of grammatical coreference
 - **coref_text.rf** – identifier (or a list of identifiers) of the antecedent(s) in the sense of textual coreference
 - **coref_special** – special types of coreference:
 - 1. **segm** – coreference with a sequence of preceding sentences (further underspecified)
 - 2. **exoph** – antecedent not present in the text at all
- **associative links ('bridging')** – to capture associative anaphora
 - Informal – types: SET, PART, FUNCT, CONTRAST, REST

Notational devices for coreferential links in PDT

- arrows from the anaphor to the antecedent(s)
- different colours of the arrows according to the type of coreference
- special devices: an exophora, a segment
- an annotator-friendly special module within the TRED editor

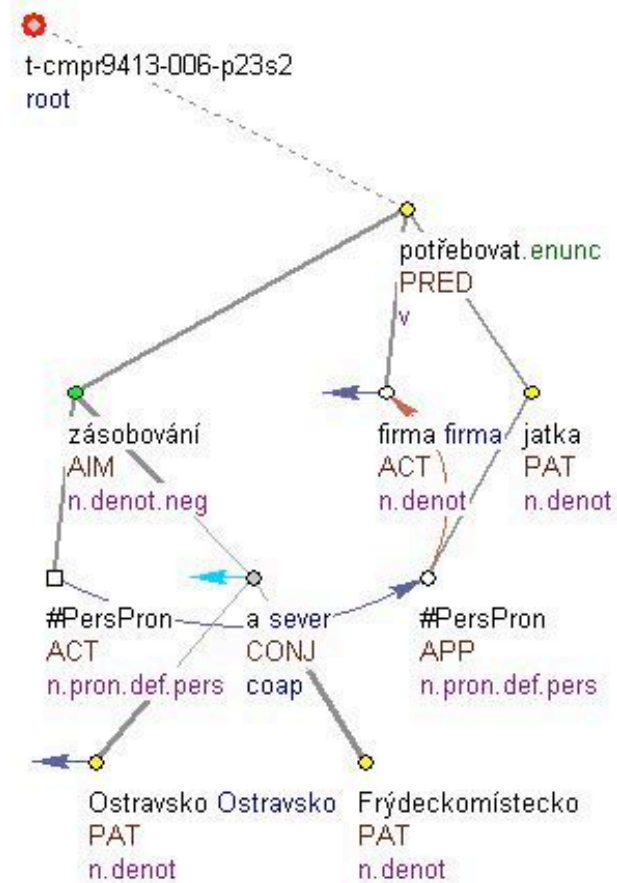
Example

- Cítí, že jen tak uskuteční svou představu.
- **[They]** feel that only in-this-way **[they]** will-realize **their** idea.
- #PersPron-ACT cítí, že jen tak #PersPron-ACT uskuteční #PersPron-APP představu.



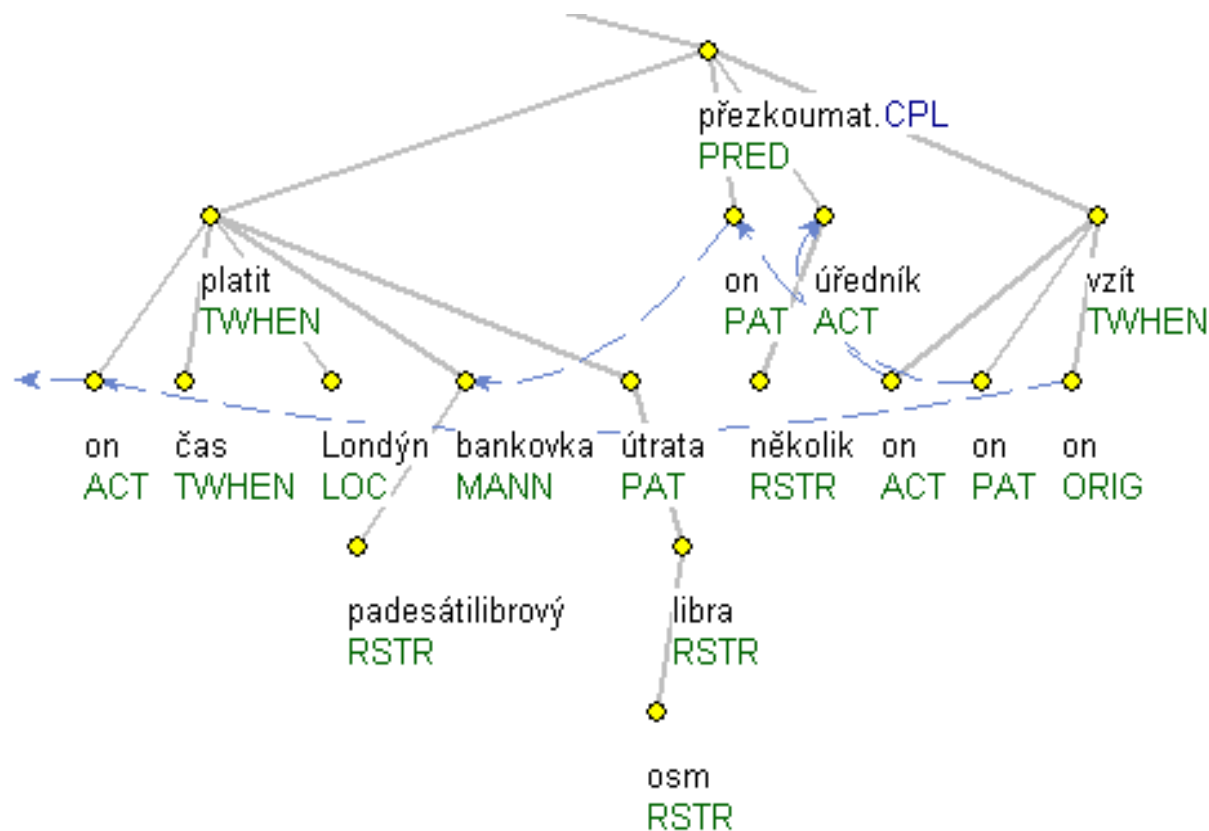
Example

- Pro zásobování Ostravska a Frýdeckomístecka firma potřebuje svá jatka.
- For [**their**] supply (of) Ostravsko and Frýdeckomístecko **the-firm** needs **their** slaughterhouse.
- Pro #PersPron-ACT zásobování Ostravska a Frýdeckomístecka firma-ACT potřebuje #PersPron-APP jatka.

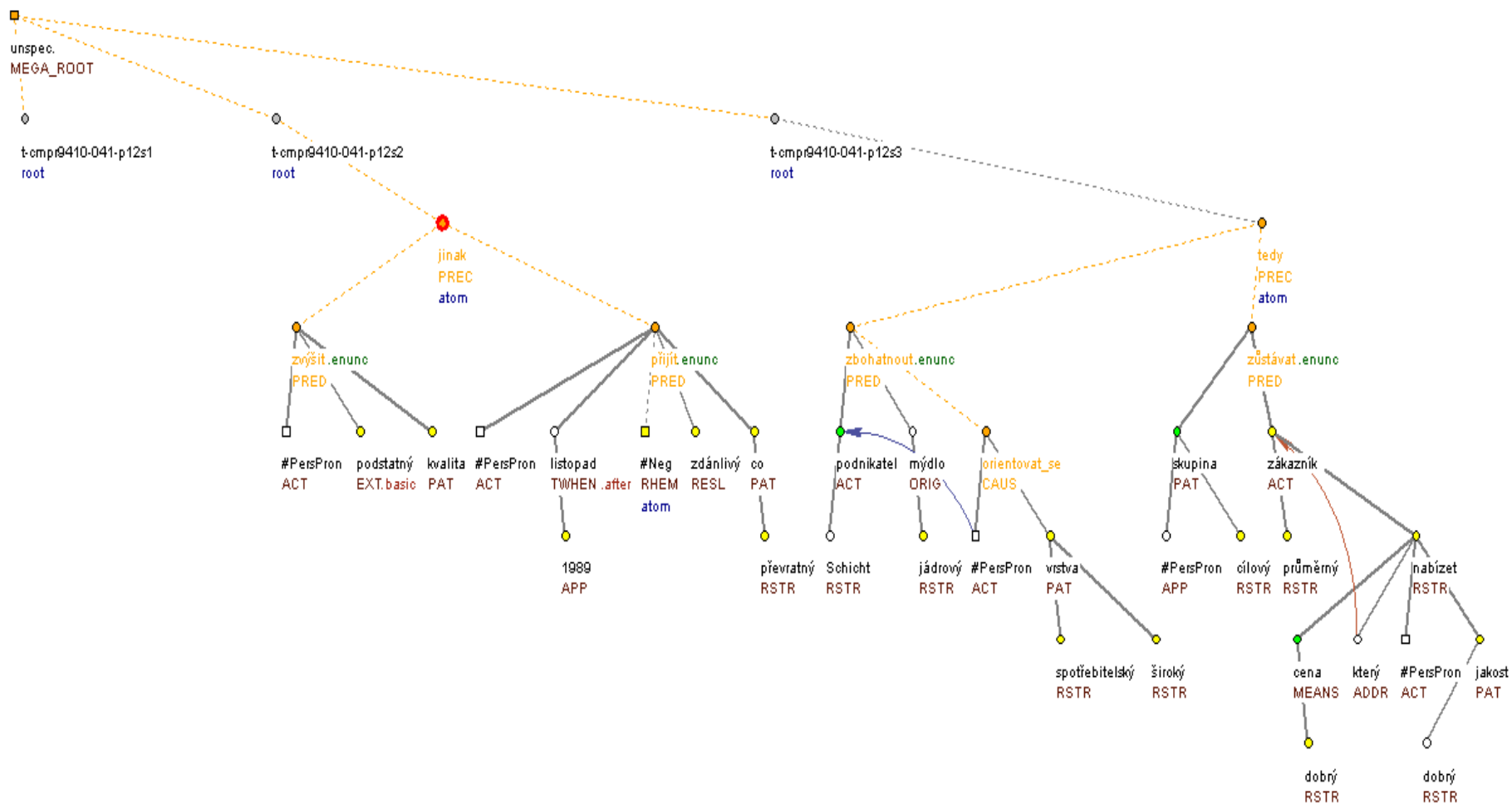


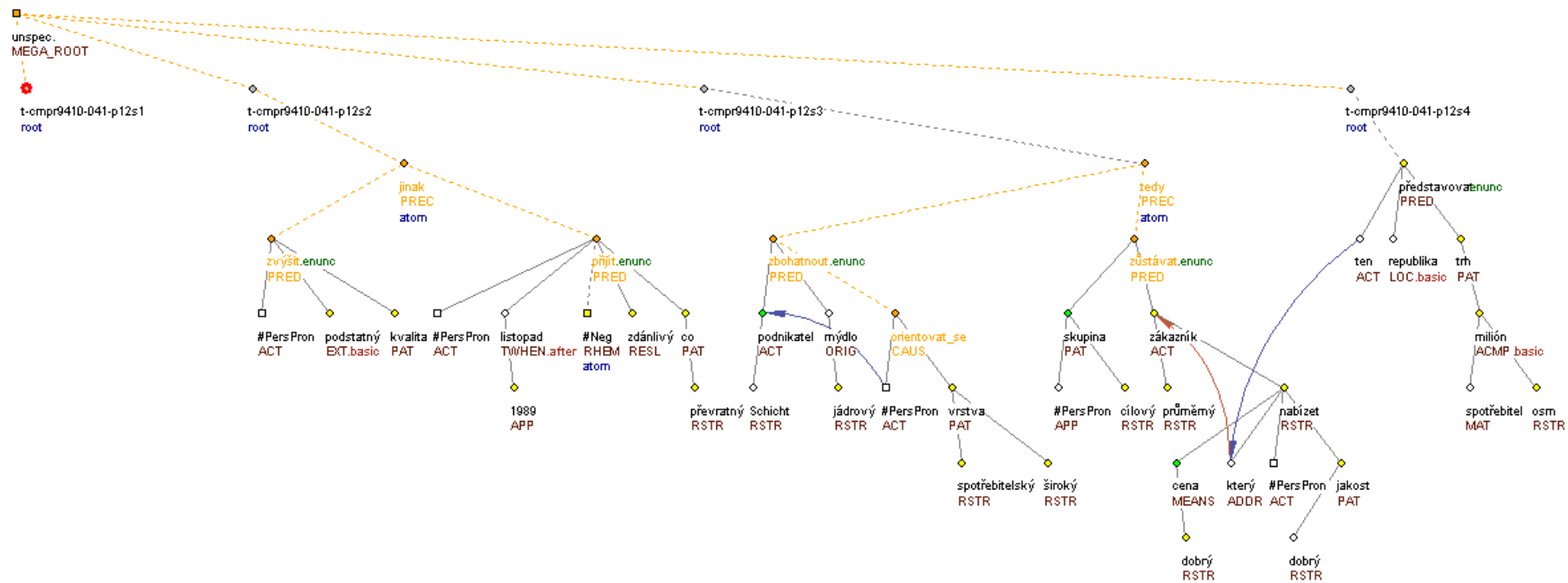
Example

- Když před časem platil v Londýně padesátilibrovou bankovkou útratu osmi liber, přezkoumalo ji několik úředníků, než ji od něj vzali.
- When time-ago **he** paid in-London with-50-pound *banknote* expenditure of-8 pounds several clerks checked *it* before they took *it* from **him**.
- Když před časem #PersPron-ACT platil v Londýně padesátilibrovou bankovkou útratu osmi liber, přezkoumalo #PersPron několik úředníků, než #PersPron-PAT od #PersPron-ORIG #PersPron-ACT vzali.



Lit.: *(When) time-ago he paid in-London with-50-pound banknote expenditure of-8 pounds, checked it several clerks (before) they took it from-him.*





Statistics: volume of data

number of annotated documents (i.e. the whole PDT 2.0 t-layer data)	3 165
number of sentences/t-trees	49 431
number of t-nodes	724 396
total number of co-referring t- nodes	46 242 (6.3% of all)

Statistics: types of coreference

grammatical coreference	23 252 (50.3%)
textual coreference	22 368 (48.4%)
special types	
segm	505 (1.1%)
exoph	120 (0.2%)

Statistics: t-lemmas with anaphors (1)

most frequent t-lemmas with grammatical coreference	
1. který	7 435 (32% of all grammatical)
2. #Cor	5 907 (25%)
3. #PersPron	4 419 (19%)
4. #QCor	2 472 (10%)
5. #Rcp	1 114 (4.7%)
6. co	575 (2.5%)
7. kde	555 (2.3%)
...	

Statistics: t-lemmas with anaphors (2)

most frequent t-lemmas with textual coreference	
1. #PersPron	18 622 (83%)
2. ten	3 733 (16.7%)
...	

Statistics: expressed vs. restored

grammatical coreference	
anaphors expressed in the surface shape	13 783 (59.3%)
restore anaphor nodes	9 469 (40.7%)
textual coreference	
anaphors expressed in the surface shape	11 131 (49.7%)
restored anaphor nodes	11 237 (50.3%)

Anaphora and bridging

- Initiated full manual **annotation of anaphora and bridging relations** in PDT
 - manual, tool, preparatory and since September 2008 full annotations, some computer experiments on automatic anaphora resolution (Anja Nedoluzhko, Giang Linh Nguy)
 - annotated on tectogrammatical trees - „all in one“
 - „our“ discourse should be added to it also (?)
- wiki pages <https://wiki.ufal.ms.mff.cuni.cz/projekt-anotace-diskurzu>

Steps beyond: segm(ent)

The boundaries of the (relevant) segment are not quite clear:

*The only reason for me to stay in America is money. [...] In America, I rent a house every year and at the end of the season I rush home. I have friends here, we go fishing, we play tennis, we visit each other. I often visit my parents in Martin. I am simply at home here. [...] In Canada **this** is totally different.*

Steps beyond: exoph(ora)

Border-line between exophora and other types of coreferential relations:

→ coreference to an unspecified element:

*A well-known native of Pardubice, Roman M. [...] had drunk himself to death after he found out that he was born in Hradec Králové. [...] The birth of children from Pardubice in Hradec Králové periodically happens. Once in every two years **[they]** brought them here, said the nurse at the obstetric clinic of the Hradec hospital.*

→ coreference to a segment („inferential“ type):

*Sad people write bright merry books and merry people write sad [ones]. One has to balance **it** somehow.*

Pronoun with other than referential function

- Intensifying function – particle *to* (*ten*):
 - *Boy, is it raining! Lit. [that] but it-rains! = meaning: it rains very much.*
- Conceptually „empty“ occurrences:
 - *As I have imagined for a long time her trip abroad, to Spain or Greece, where [lit.] it draws her.*
- Phrasemes
 - *Lit. That you-have hard, this young person's father has connections.*

Open questions (1)

Coreferential link leads to the root

× antecedent is a part of sentence:

*When Jiří Krupička sent me the manuscript of his Renaissance of Reason, which has been published now in the publishing house Český spisovatel, and I looked into it for the first time, not only my knees but also my heart trembled. And **this** [happened] for several reasons.*

Open questions (2)

With a coreferential chain, all links are established:

*The agreement of course has not solved anything – it only deepened the feeling in the **protestants** that London leaves **them** in the lurch. Today this feeling, that **[they]** are only a burden for Great Britain, which **[they]** do not know how to deal with, has strengthened in Ulster protestants.*

Open questions (3)

Nodes are reconstructed also with nominalizations:

*It [=the word] has a strong emotive **colouring** and it occurs especially in discourse of young people.*

colouring → Gen.ACT

→ Gen.PAT → *on*.PAT → *slovo* [word]

Open questions (4)

- Abstract nouns
 - The *unemployment* rate should develop in another way than in standard economics. ... [...] the increase of the *unemployment* rate in 1991-1993 is lower than ... The progressing privatization and restructuralization will cause the raise of the *unemployment* rate from 3.5% to 5% at the end of the next year.
- Should be annotated at this stage?
- If yes – which type - 0 or NR (at present – NR)?
- The boundary between abstract and concrete – difficult:
profit

Process of annotation interannotators' agreement

- Beginning: 10.2008
- 10.-12.2008: 3
annotators 7000 vět
- 1.2009 -: 2
annotators
- 1000 sentences per
month
- -> 1/2 PDT by the end
of 2009

