

Corpus Annotation – Sentence and Discourse

4. From the structure of the sentence to discourse patterning

Aim

- (1) To document **which aspects** of the **discourse** structure can be **discovered** and elucidated from the **sentence annotation** on an underlying syntactic layer including **Topic-Focus Articulation** and an establishment of basic **coreference** links in the **Prague Dependency Treebank**.
- (2) **Discourse annotation in Penn Discourse treebank and in PDT.**

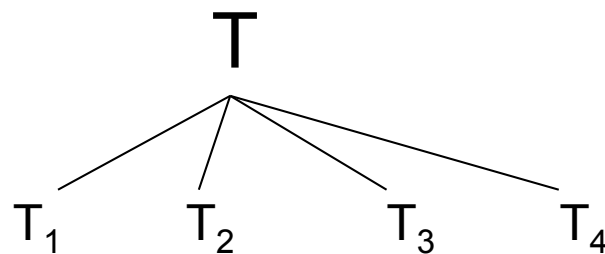
A. Relationships between sentence and discourse patterning

- Henri Weil (1844): ‘progressions’ of ideas
- (i) parallel (ii) progression

$I_1 - I_2$
 $I_1 - I_3$
 $I_1 - I_4$

$I_1 - I_2$
 $I_2 - I_3$
 $I_3 - I_4$

- František Daneš (1970): thematic progressions
- (i), (ii), (iii)



B. Stock of shared knowledge

- SSH: a structured whole
- Hierarchy of activation of the SSK elements – a partial ordering
- Heuristic rules for the assignment of degrees of activation
- Implementation of the rules and visualization of the results

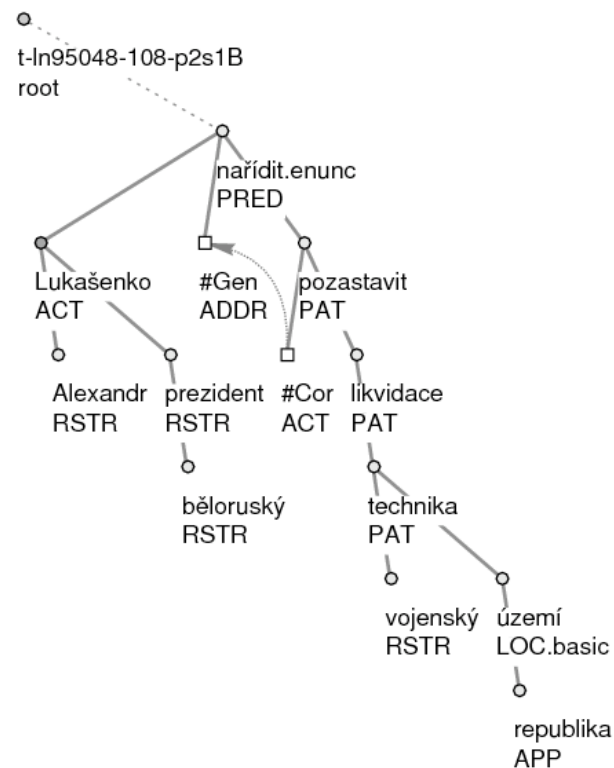
Our hypothesis

- Discourse structure analysis:
 - **Hypothesis:** A finite mechanism exists that enables the addressee to identify the referents on the basis of a partial ordering of the elements in the stock of knowledge shared by the speaker and the addressees (according to the speaker's assumption), based on the degrees of activation of referents.

Prague Dependency Treebank:

Three layers of annotation

- morphological
- syntactical
- **tectogrammatical**



Běloruský prezident Alexandr Lukašenko nařídil pozastavit likvidaci vojenské techniky na území republiky.

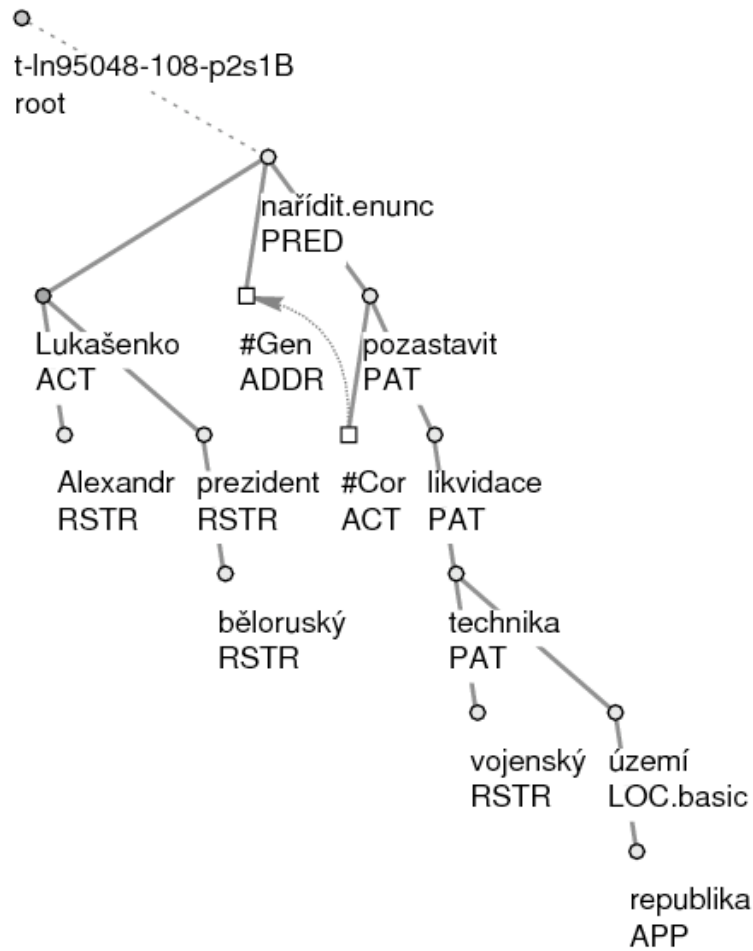
lit. Belorussian president Alexandr Lukashenko ordered to-stop the-liquidation of-military technology on-the- territory of-the-Republic.

Degrees of activation

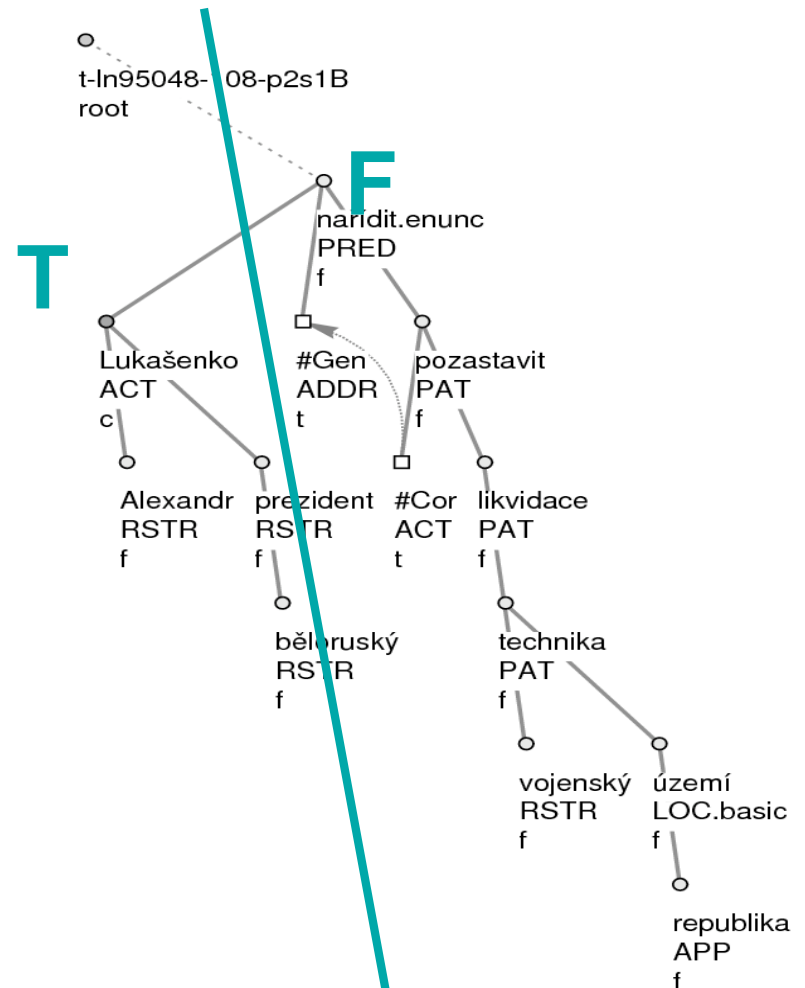
- degrees of activation of the elements of SSK (Hajičová, 1993)
- heuristics based on
 - the position of the items in question in the topic or in the focus of the sentence
 - the means of expression (noun, pronoun)
 - the previous state of the activation

Prague Dependency Treebank: Topic-Focus articulation

- bipartition of the sentence – based on the **TFA values**
 - **t** - contextually bound non-contrastive
 - **c** – contextually bound – contrastive
 - **f** – contextually non-bound
- algorithm – implemented, tested



Běloruský prezident Alexandr Lukašenko nařídil pozastavit likvidaci vojenské techniky na území republiky.

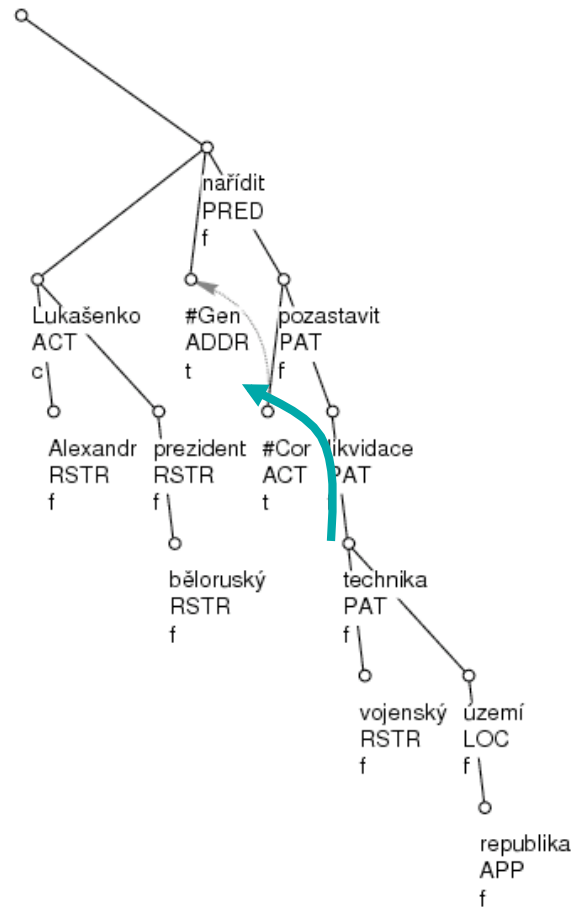


Běloruský prezident Alexandr Lukašenko nařídil pozastavit likvidaci vojenské techniky na území republiky.

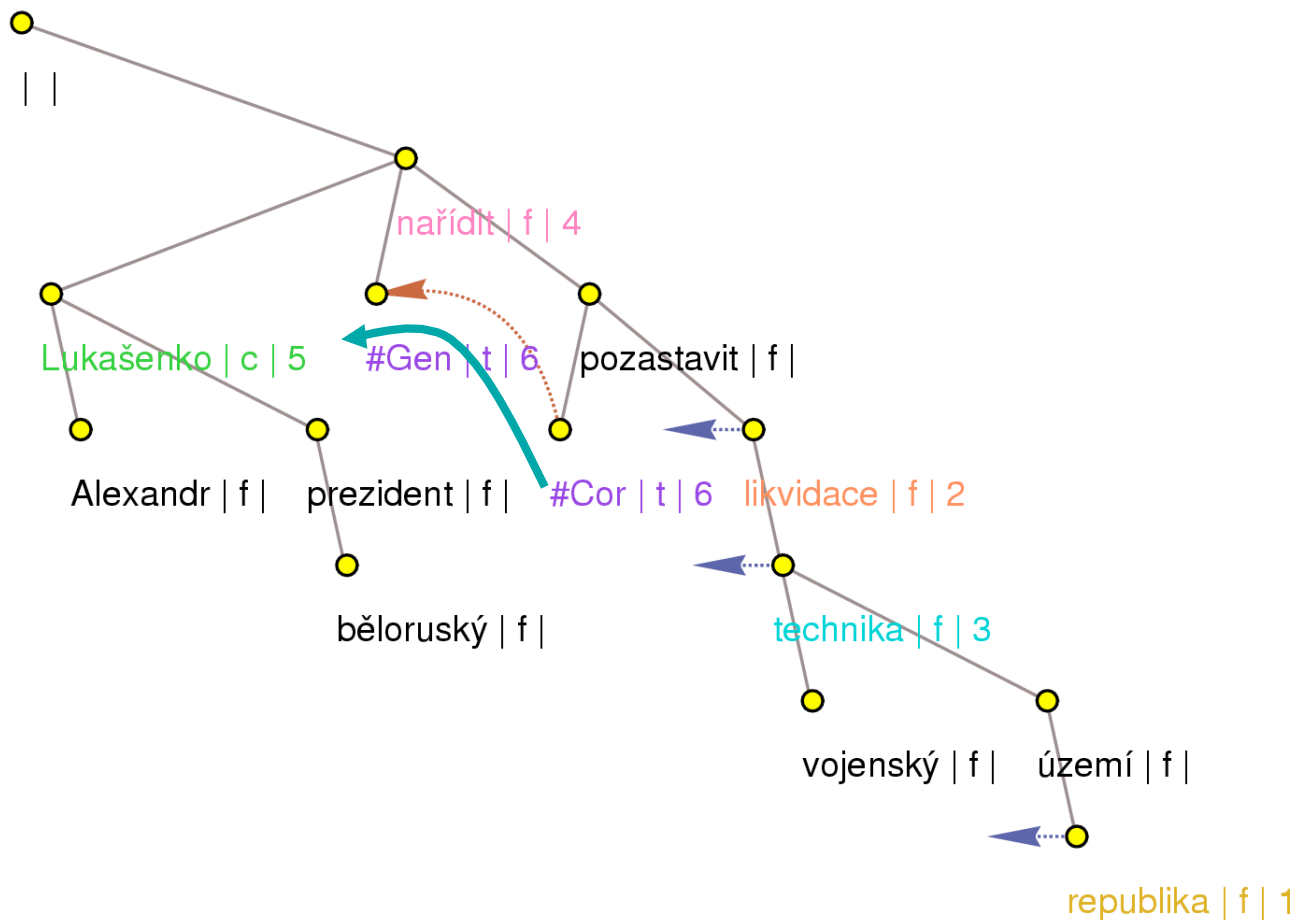
Prague Dependency Treebank: Coreference

PDT 2.0

- textual coreference
 - relations of control
 - relative pronouns (wh-words)
 - reflexive pronouns
- in addition
 - anaphors expressed by pronouns (or reconstructed, pro-drop!)



Běloruský prezident Alexandr Lukašenko nařídil pozastavit likvidaci vojenské techniky na území republiky.
lit. Belorussian president Alexandr Lukashenko ordered to- stop the-liquidation of-military technology on-the-territory of-the-Republic.



Běloruský prezident Alexandr Lukašenko nařídil <#Gen> pozastavit <#Cor> likvidaci vojenské techniky na území republiky.

lit. Belorussian president Alexandr Lukashenko ordered to-stop the-liquidation of-military technology on-the-territory of-the-Republic.

1. **Bělorusko**₁: zastavení **likvidace**₂ **arzenálu**₃. (E: Belorussia: stopping the liquidation of arsenal.)
2. ...
3. Běloruský prezident Alexandr **Lukašenko**₅ **nařídil**₄ < #Gen >₆ pozastavit < #Cor >₆ **likvidaci**₂ vojenské **techniky**₃ na území **republiky**₁. (E: Belorussian president Alexandr Lukashenko ordered to stop the liquidation of military technology on the territory of the Republic.)
4. ...
5. ...
6. ...
7. Agentura Interfax soudí, že **prezident**₅ měl na mysli přání východoevropských **zemí**₇ **vstoupit**₈ < #Cor >₇ do Severoatlantické **aliance**₉, **což**₈ by pro **Bělorusko**₁ znamenalo bezprostřední sousedství s **NATO**₉. (E: The agency Interfax assumes that the president had on mind the wish of East-European countries to enter into the North-Atlantic alliance, which would for Belorussia mean immediate neighborhood with NATO.)
8. **Lukašenko**₅ také řekl, že < #PersPron >₅ má pochybnosti < #QCor >₅ o dosud deklarovaném neutrálním statusu **Běloruska**₁, uvedl, < #PersPron >₅ že je pro nový systém národní bezpečnosti, a oznámil, že < #PersPron >₅ ustavil "pracovní skupinu pro vytvoření vojenské doktríny **Běloruska**₁". (E: Lukashenko also said that he has doubts about the hitherto declared neutral status of Belorussia, he stated that he is for a new system of national security, and announced that he put-together "a working group for creation of military doctrine of Belorussia".)
9. Přitom ujistil < #PersPron >₅, že v souladu s republikovou legislativou "žádný běloruský voják nebude bojovat za hranicemi **Běloruska**₁". (E: At the same time he assured that in accordance with Republic's legislation "no Belorussian soldier will

The salience algorithm

- to capture the dynamic character of SSK
- processing sentence by sentence
- salience degree $dg^n(x)$ of an item x represented by the referent r after the n -th sentence of a document is uttered

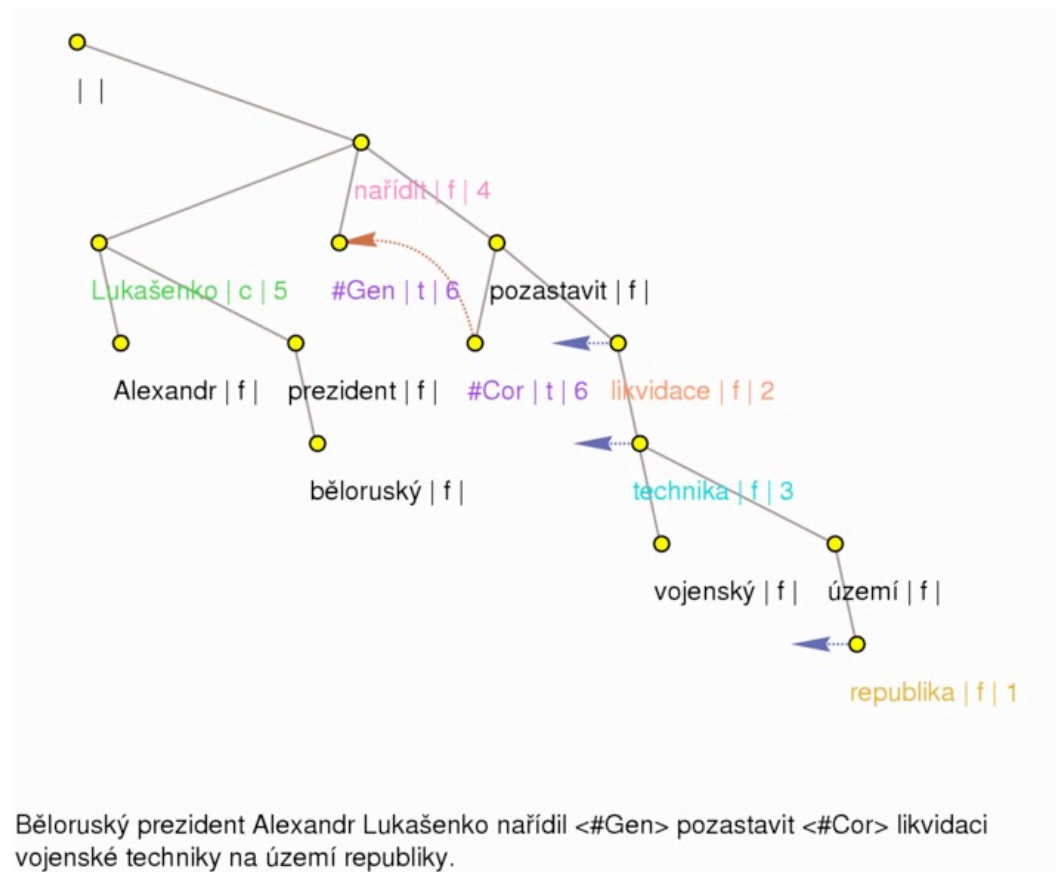
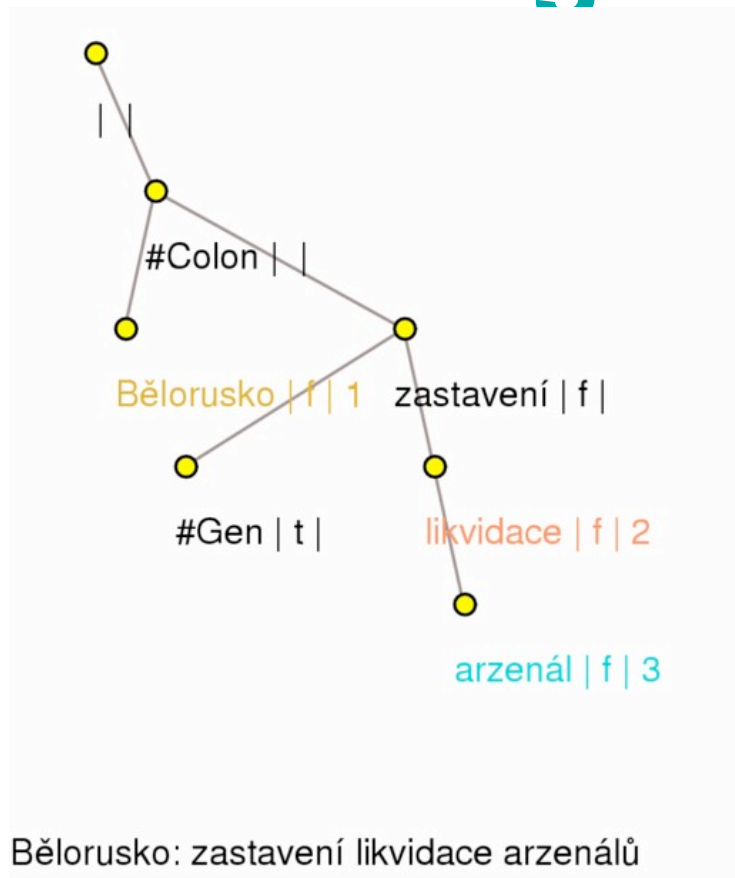
The salience algorithm - heuristics

$dg^n(x)$ to be read as 'an item x represented by the referent r has the salience degree $dg^n(x)$ after the n -th sentence of a document is uttered, i.e. salience degree of the item is modified after each sentence starting with sentence in which the item has appeared firstly':

1. $dg^n(x) = -1$ if r carries TFA value t or c in the n -th sentence.
2. $dg^n(x) = 0$ if r carries TFA value f in the n -th sentence.
3. $dg^n(x) = dg^{n-1}(x) - 2$ if r is not included in the n -th sentence and has been mentioned in the Focus of the last (not necessarily immediately) preceding sentence ($(n-1)$ -th through 1-st sentence).
4. $dg^n(x) = dg^{n-1}(x) - 1$ if r is not included in the n -th sentence and has been mentioned in the Topic of the last (not necessarily immediately) preceding sentence ($(n-1)$ -th through 1-st sentence).

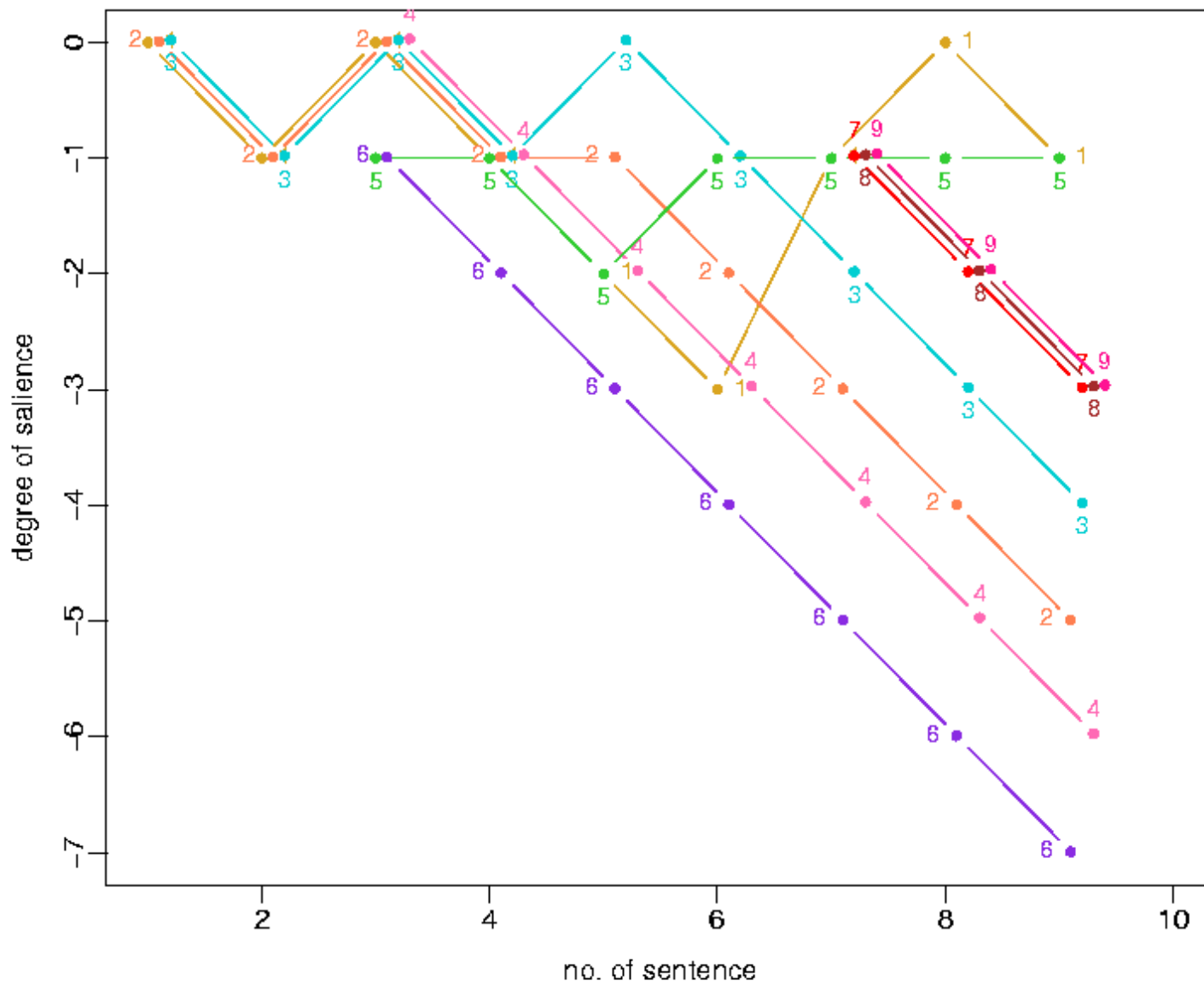
1. **Bělorusko**₁: zastavení **likvidace**₂ **arzenálů**₃. (E: Belorussia: stopping the liquidation of arsenal.)
2. ...
3. Běloruský prezident Alexandr **Lukašenko**₅ **nařídil**₄ < #Gen >₆ pozastavit < #Cor >₆ **likvidaci**₂ vojenské **techniky**₃ na území **republiky**₁. (E: Belorussian president Alexandr Lukashenko ordered to stop the liquidation of military technology on the territory of the Republic.)
4. ...
5. ...
6. ...
7. Agentura Interfax soudí, že **prezident**₅ měl na mysli přání východoevropských **zemí**₇ **vstoupit**₈ < #Cor >₇ do Severoatlantické **aliance**₉, **což**₈ by pro **Bělorusko**₁ znamenalo bezprostřední sousedství s **NATO**₉. (E: The agency Interfax assumes that the president had on mind the wish of East-European countries to enter into the North-Atlantic alliance, which would for Belorussia mean immediate neighborhood with NATO.)
8. **Lukašenko**₅ také řekl, že < #PersPron >₅ má pochybnosti < #QCor >₅ o dosud deklarovaném neutrálním statusu **Běloruska**₁, uvedl, < #PersPron >₅ že je pro nový systém národní bezpečnosti, a oznámil, že < #PersPron >₅ ustavil "pracovní skupinu pro vytvoření vojenské doktríny **Běloruska**₁". (E: Lukashenko also said that he has doubts about the hitherto declared neutral status of Belorussia, he stated that he is for a new system of national security, and announced that he put-together "a working group for creation of military doctrine of Belorussia".)
9. Přitom ujistil < #PersPron >₅, že v souladu s republikovou legislativou "žádný běloruský voják nebude bojovat za hranicemi **Běloruska**₁". (E: At the same time he assured that in accordance with Republic's legislation "no Belorussian soldier will

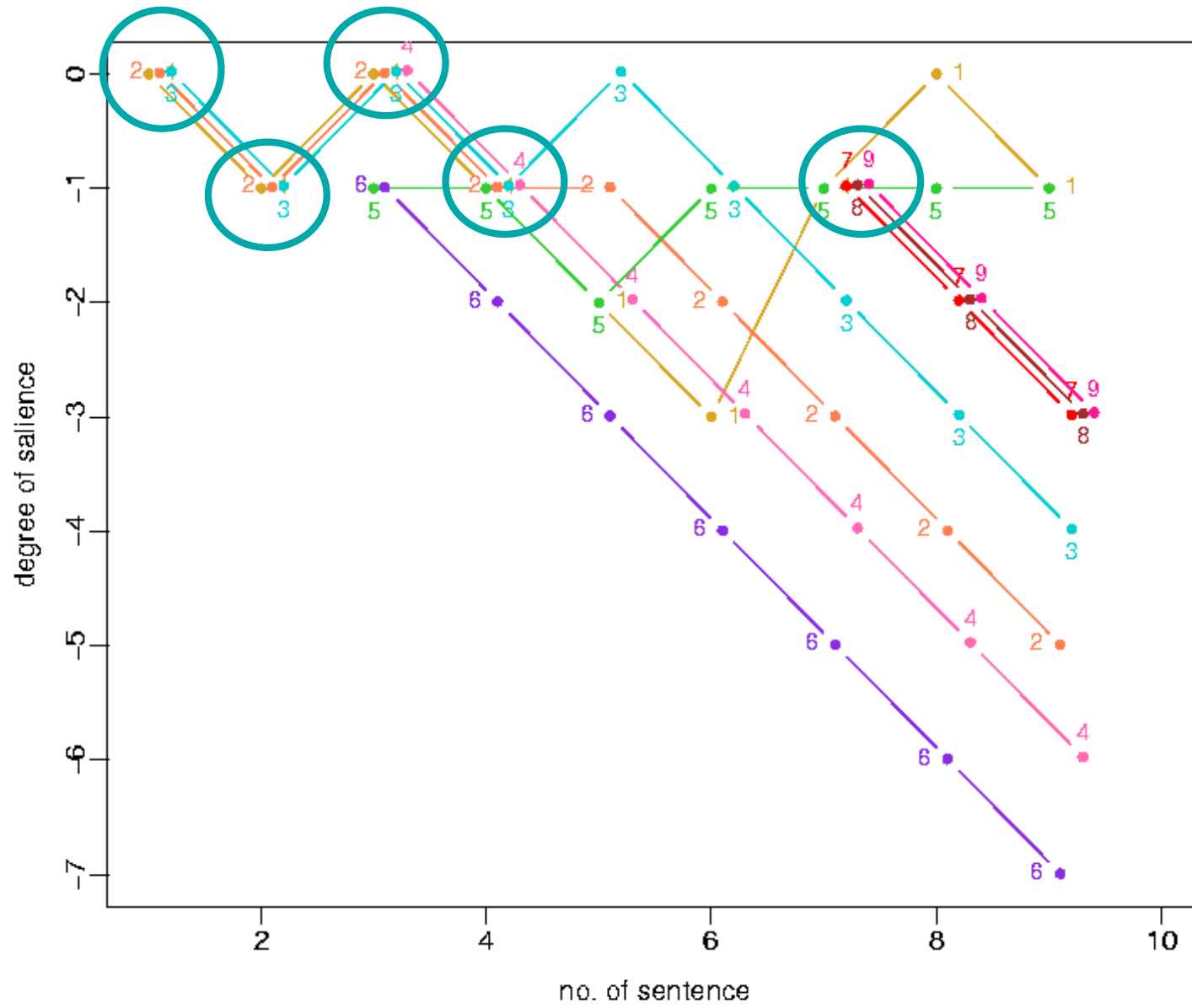
Tectogrammatical trees

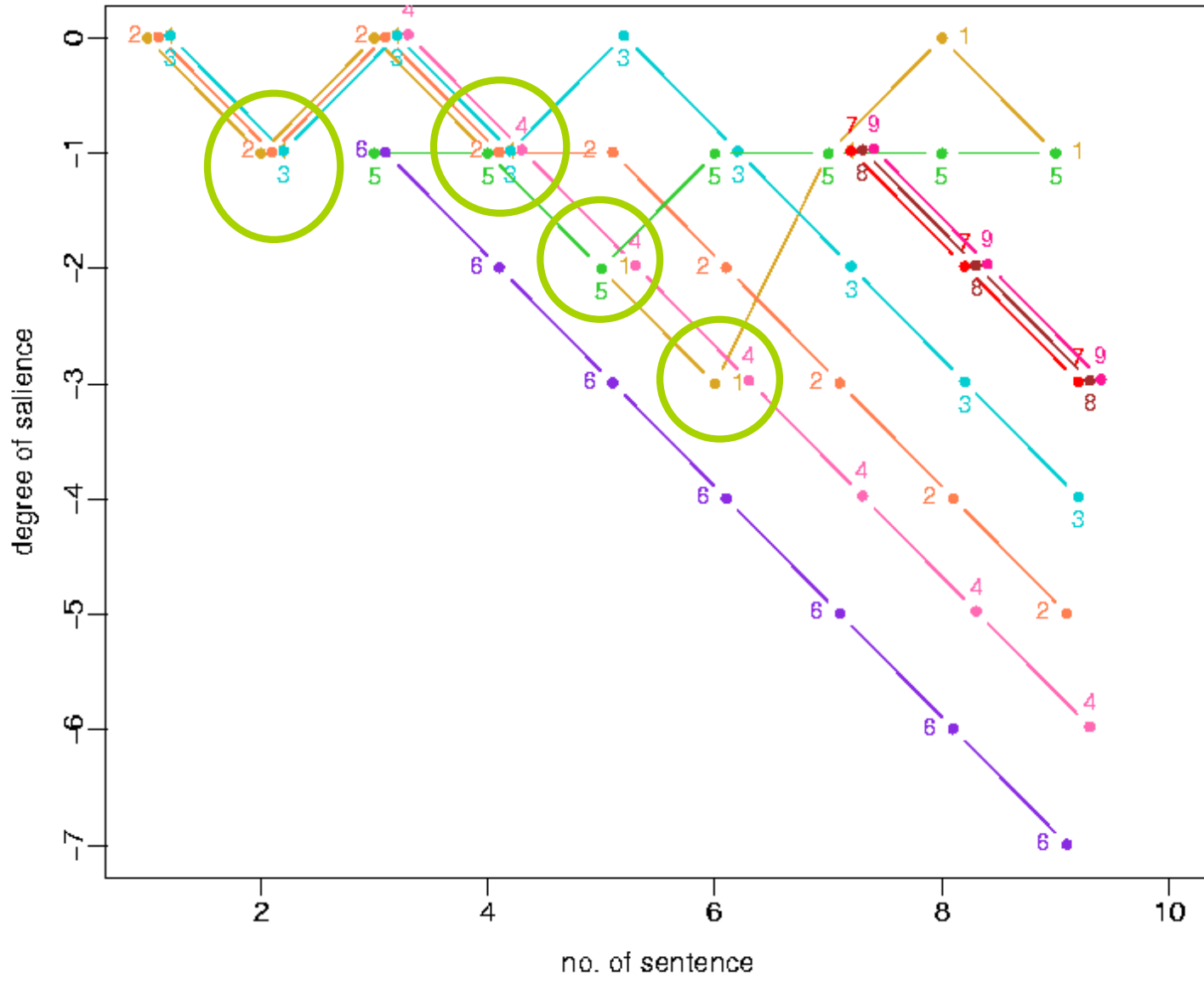


Sample document – coreference chains

- [1: Belarussia] Bělorusko/f/F/1 republika/f/F/3 Bělorusko/c/F/7 Bělorusko/t/F/8 Bělorusko/f/F/8 Bělorusko/t/F/9
- [2: liquidation] likvidace/f/F/1 likvidace/f/F/3 opatření/t/T/5
- [3: arsenal] arzenál/f/F/1 technika/f/F/3 a[tank/f/F/5 letadlo/f/F/5 transportér/f/F/5 vozidlo/f/F/5
- [4: to order] nařídít/f/F/3 ten/t/T/4
- [5: Lukashenko] Lukašenko/c/T/3 #PersPron/t/T/4 Lukašenko/t/T/6 prezident/t/F/7 Lukašenko/t/T/8 #PersPron/t/F/8 #QCor/t/F/8 #PersPron/t/F/8 #PersPron/t/F/8 #PersPron/t/T/9
- [6] #Gen/t/T/3 #Cor/t/F/3
- [7: country] země/f/F/7 #Cor/t/T/7
- [8: to enter] vstoupit/f/F/7 co/t/F/7
- [9: NATO] aliance/f/F/7 NATO/t/F/7







How useful it is

Three tentative hypotheses:

1. The changes of activation during the piece of discourse (i.e. a document) indicate a division of the document into “informative” segments.
2. On the basis of the prominent activation of some elements of the SSK in these segments the ‘topics’ of segments (and, consequently, of the documents) can be specified.
3. The vertical lines in the scheme may help decide about pronominal reference assignment.

Example

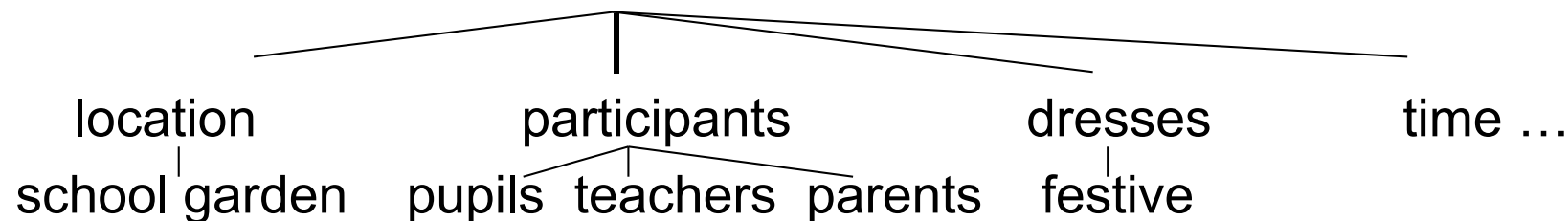
- (1) The school garden was full of CHILDREN.
- (2) They talked NOISILY,
- (3) but the teachers didn't REPROVE them
- (4) because they were so EXCITED.
- (5) Outside, PARENTS were waiting.
- (6) One of them, a father, stood in front of a MICROPHONE.
- (7) The pupils got CALM
- (8) and their teachers lined them UP.
- (9) Both teachers and pupils were in a festive MOOD.
- (10) The teachers were SERIOUS.
- (11) In fact, ALL adults in the garden were serious.
- (12) They were dressed in evening DRESSES.
- (13) As for the pupils, they had school UNIFORMS.
- (14) neatly washed and PRESSED.
- (15) The smallest even had snow-white COLLARS.
- (16) One of the parents approached the biggest BOY
- (17) and ASKED him:
- (18) "Is it allowed to sit down on the ground?"

- Segment topics:
 - children
 - teachers = adults
 - parents = adults
 - Sentence 4: they?
 - associated items
 - boy, the smallest, ...
 - father, ...

Pronominal reference

- “*they*” – if the degree of activation is almost equal → uncertainty
- if the item is too far – some specific means: e.g. *as for ...*
- what is the document about? “frame”

School festival



Play the Coreference

Getting more data

- an innovative method to get annotated data:
an internet game
- players mark words that refer to the same entity
- no knowledge of linguistic terms regarding coreference required from the players
- outcome – data for machine learning approaches to automatic coreference resolution, name entity recognition

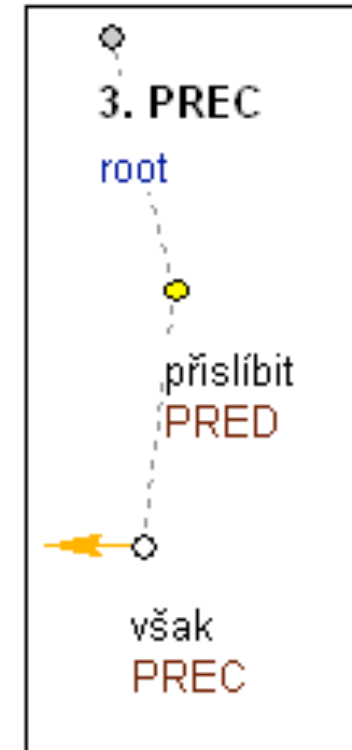
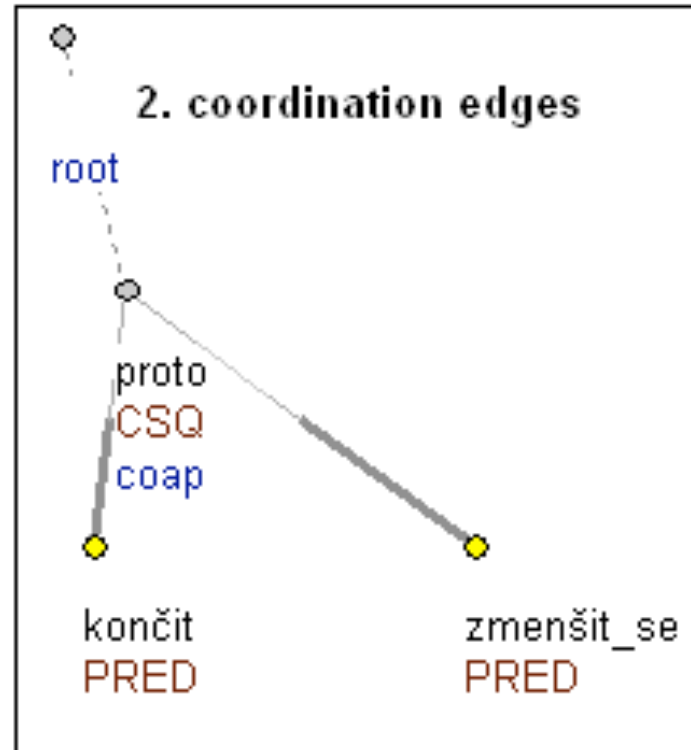
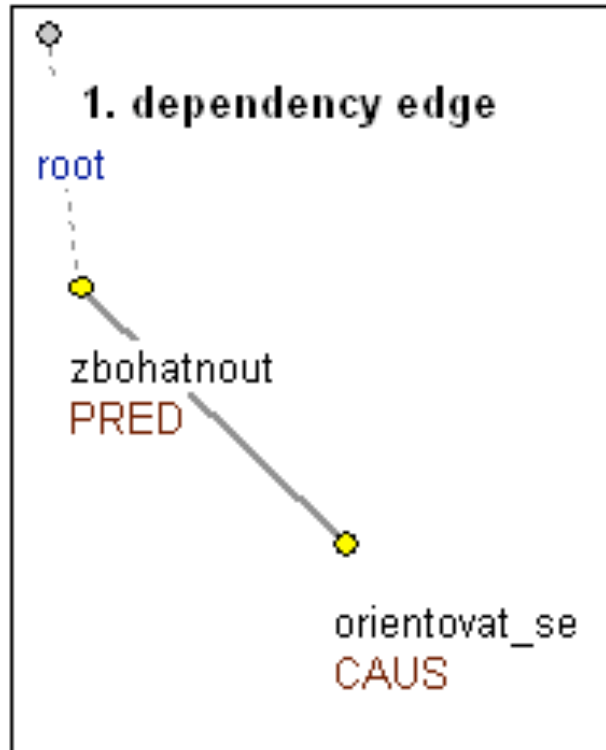
C.Annotation of discourse in different **annotation** schemes

- Penn Discourse Treebank: corpus of English texts approx. 49000 sentences, Wall Street Journal
- underlying idea: structuring of text by means of lexical items – ‘connectives’
- each connective: treated as a discourse level predicate, 2 arguments (text spans – a whole sentence or a sentence part)
- may be located in a distance or interrupted

Penn Discourse Treebank (2)

- Penn Discourse Treebank – discourse relations annotated independently on the syntactic annotation scheme of Penn Treebank
- An interesting issue: how far the Prague Dependency Treebank syntactic relations can help on the way from individual sentences to discourse

From PDT to discourse



Current status of discourse relations in PDT

Only within a single tree:

- Paratactic relations (coordination)
- Hypotactic relations (subordination) CAUS, COND, AIM, CNCS, TWHEN, LOC, DIR, MANN, ACMP, REG
- Reference to the preceding context (PREC, event. RHEM, CM)
- No annotation yet: exemplification, implicit intersentential relations

Sense tags comparison

(The changes in the original hierarchy are marked with *italics and green*.)

TEMPORAL

- asynchronous
 - precedence
 - succession

- synchronous

CONTINGENCY

- cause
 - reason
 - result

 - consequence*

 - result of the high degree of a property (RESL)*

- condition

 - hypothetical
 - general
 - unreal present
 - unreal past
 - factual present
 - factual past

 - result of the condition*

 - concession*

 - expectation
 - contra-expectation

 - purpose*

MANNER

 - way*

 - means*

 - criterion*

 - regard*

Sense tag comparison

COMPARISON

contrast
juxtaposition (CONTRA, CONTRD)
opposition (ADVS)
correction
scale
gradation (GRAD)
level (CPR)
difference (DIFF)

EXPANSION

conjunction (CONJ)
instantiation
restatement
specification
equivalence (sentential APPS)
generalization
alternative
conjunctive (CONJ or DISJ)
disjunctive (DISJ)
chosen (SUBS)
exception (RESTR)
list

DIRECTIONAL (LOCAL) FUNCTORS?

unspec.
MEGA_ROOT

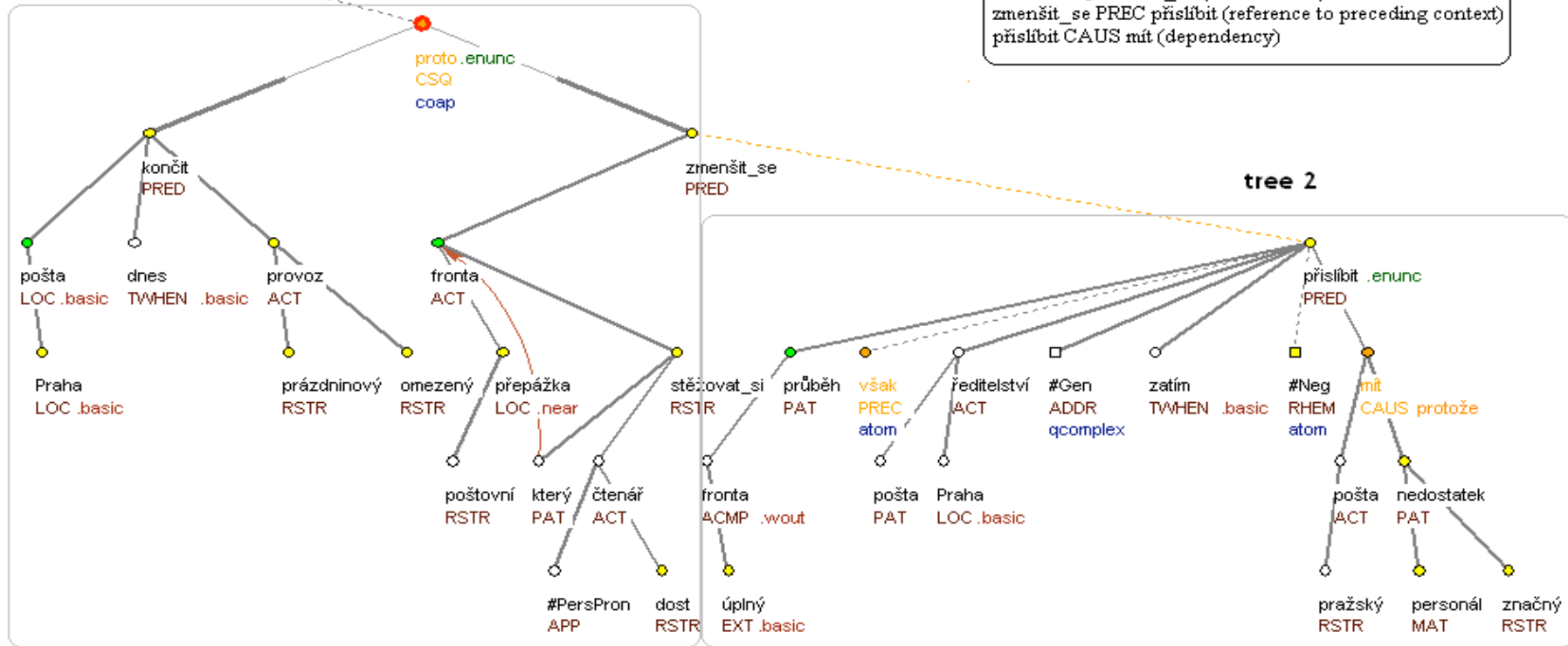
t-ln94205-47-p2s1B
root

(Lit.) tree 1: [At the post offices in Prague today, (there is) ending (PRED) the restricted holiday operation], [the queues at the counters, about which a lot of our readers have complained, should **therefore** (CSQ, coordination) shorten (PRED)].
tree 2: [An operation completely without queues, **however** (PREC), the post management in Prague for now cannot guarantee (PRED)] [**because** (hidden, CAUS) the Prague post has (CAUS, dependency) a considerable lack of staff.]

tree 1

tree 2

Discourse relations:
končit CSQ zmenšit_se (coordination)
zmenšit_se PREC přislíbit (reference to preceding context)
přislíbit CAUS mít (dependency)



- [Unit 1] discourse connective [Unit 2]
- (1) [What had you been like] before [you lost your job?]
- DC = before

Penn: Temporal, asynchronous, precedence

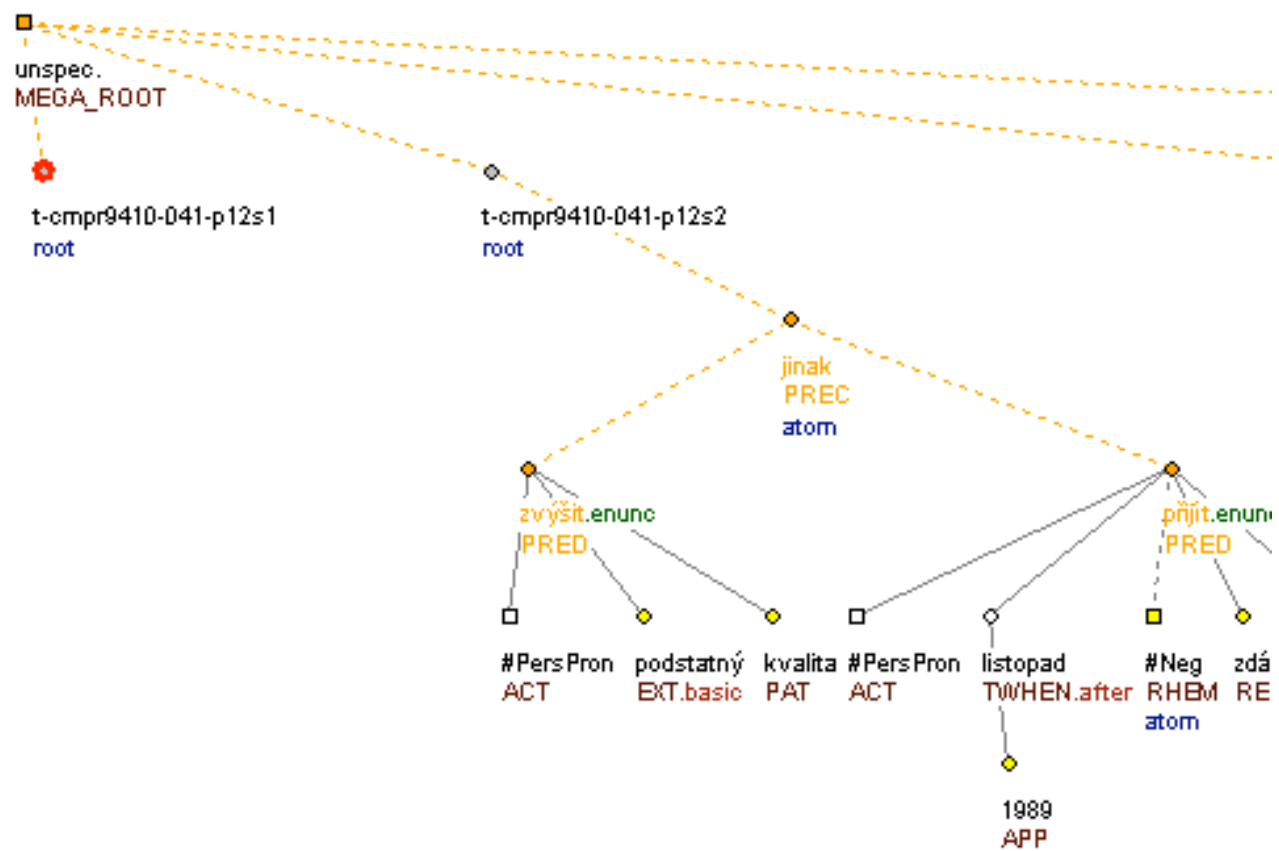
PDT: functor TWHEN, subfunctor BEFORE

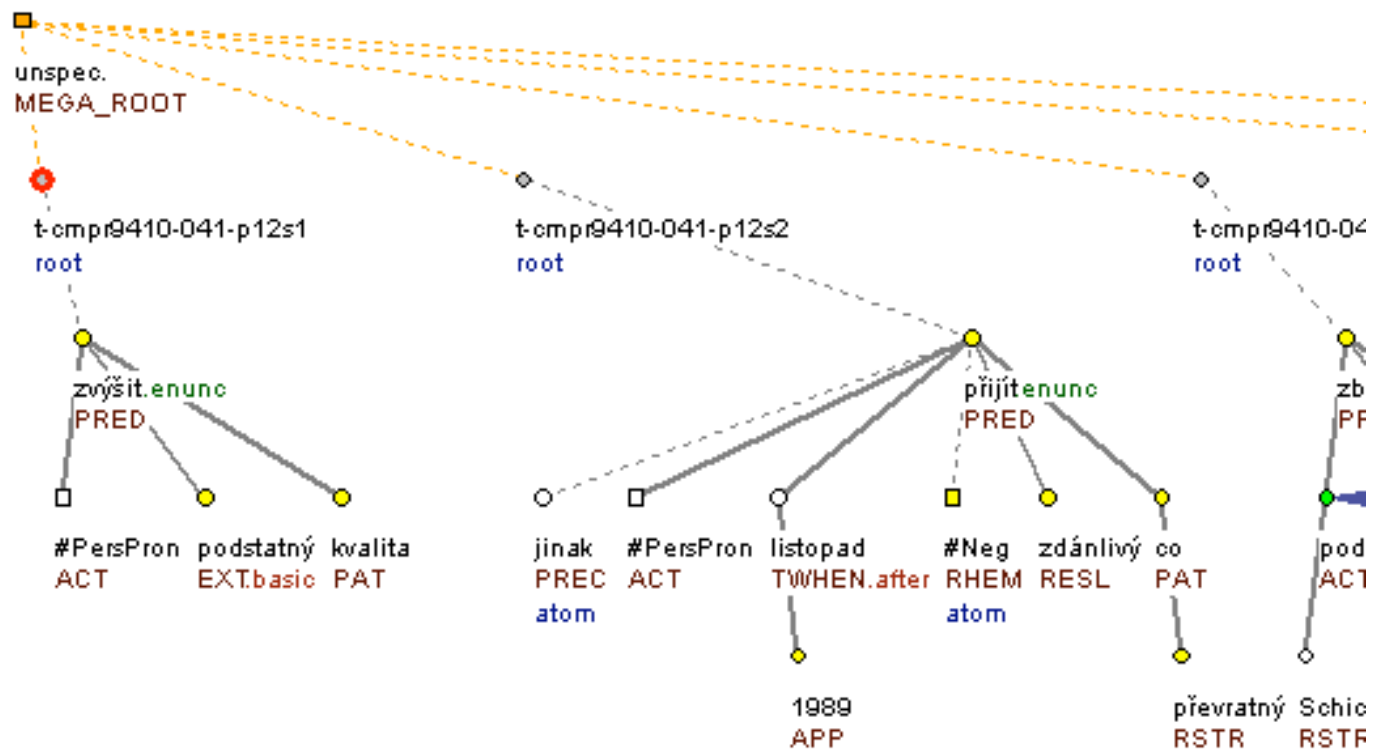
- [Either we will go to the cinema] or [we'll stay at home.]
- DC = or (disjunctive meaning)
- Penn: expansion – alternative – disjunctive
- PDT: functor DISJ

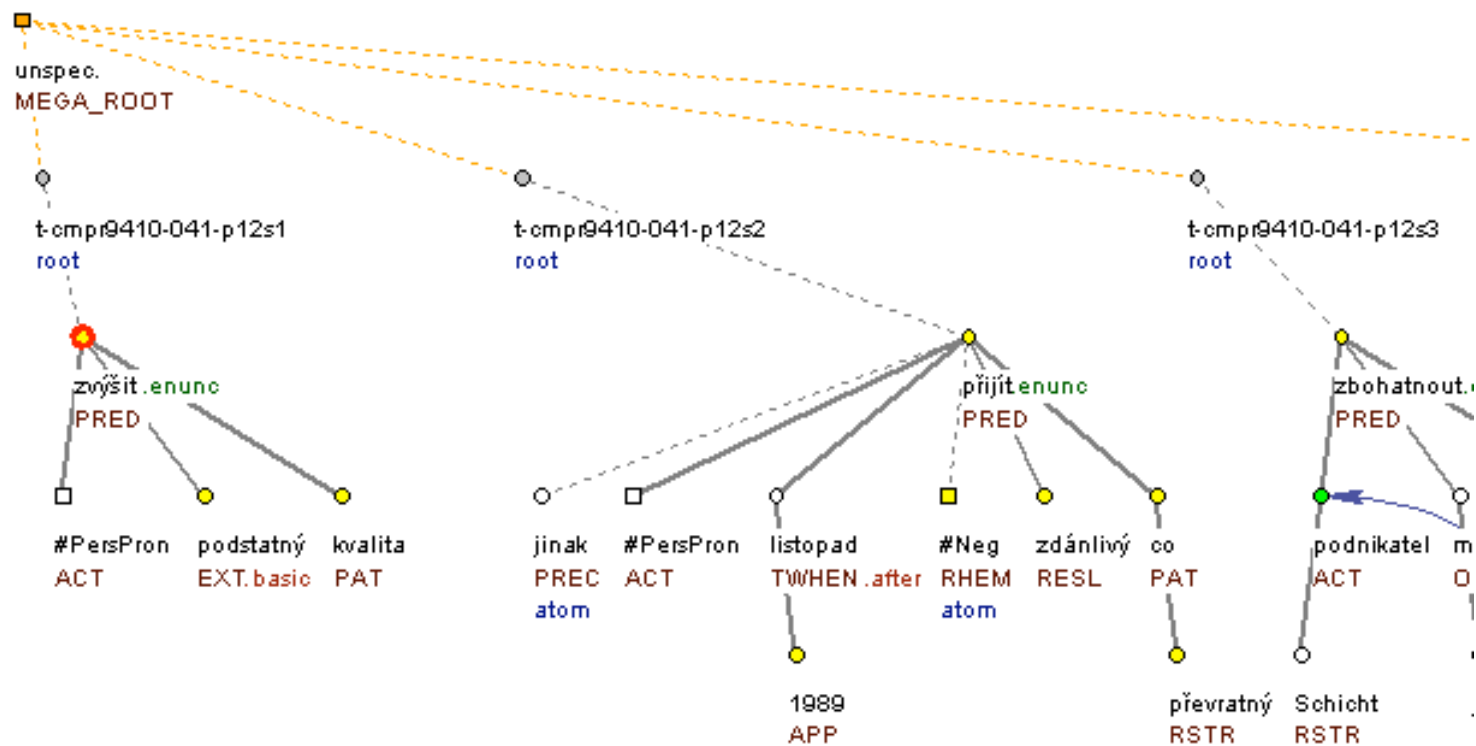
- [...] And [then he left.]
- DC = and
- Penn: expansion – conjunction
- PDT: functor PREC (no discourse semantics marked yet)

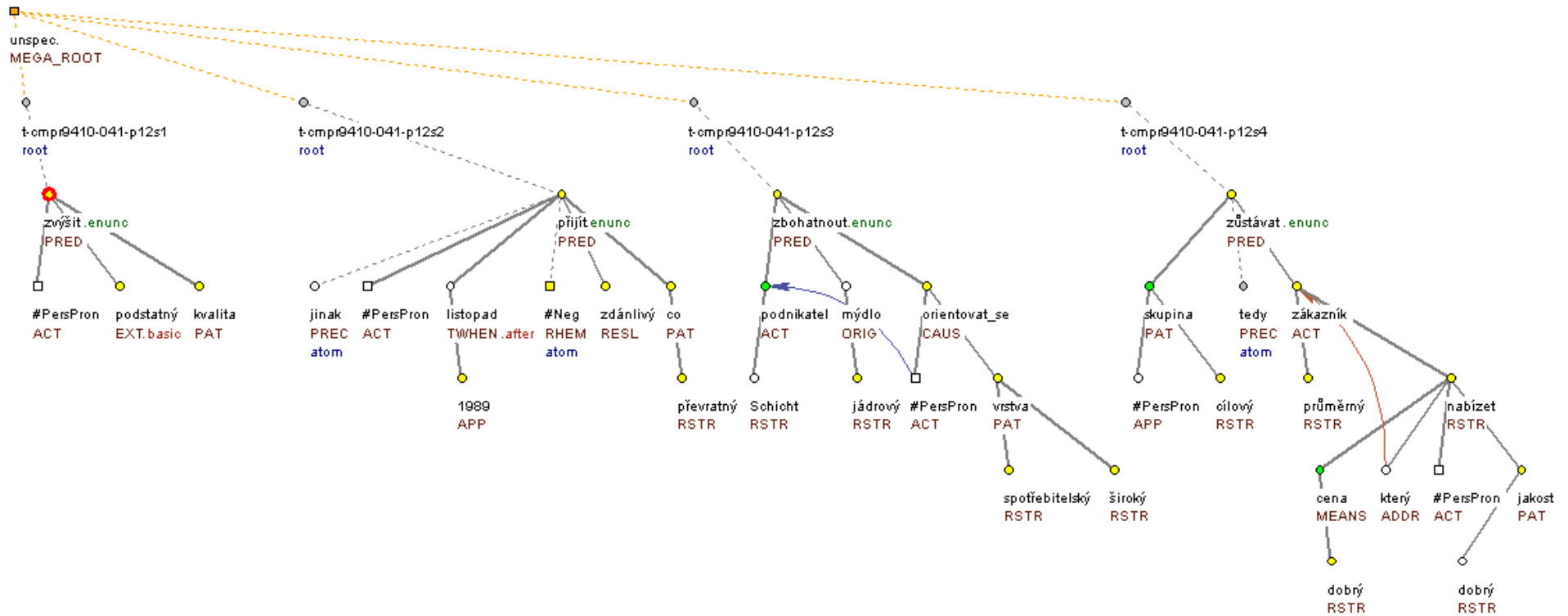
- (28) Podstatně jsme zvýšili kvalitu .
- (29) Jinak jsme po listopadu 1989 zdánlivě nepřišli s ničím převratným .
- (30) Podnikatel Schicht zbohatl na jádrovém mýdle, protože se orientoval na nejširší spotřebitelskou vrstvu .
- (31) Naší cílovou skupinou tedy zůstává průměrný zákazník , kterému za dobrou cenu nabízíme dobrou jakost .
- (32) **To** představuje v republice trh s osmi miliony spotřebiteli .

- (28) We substantially improved the quality.
- (29) Otherwise, after November 1989 we apparently have not introduced anything revolutionary.
- (30) The entrepreneur Schicht got richer by the special soap, because he has oriented on the broadest scope of consumers.
- (31) Our target group is thus an average consumer, to whom we offer a good quality for a good price.
- (32) **This** represents in our republic a market with eight million consumers.









Open questions (1)

- Annotation on a linear text or on (mega) trees?
- A questionable binarity of connectors
- Enriched TGTS's or an introduction of a special layer of discourse?
- Representation of discourse relations – a connector with two arguments or a labelled edge between two arguments?

Open questions (2)

- A clear specification of discourse relations
- Relation of subordination and coordination and its relation to discourse relations
- A more precise (and subtle?) classification of subordination relations
- The relationships between discourse relations and lexical meaning

Conclusions

- a systematic annotation of a large corpus of (segments of) continuous text(s) on several layers has an indisputable advantage
- there are, of course, many other respects in which corpus annotation schemes should go beyond the current practice
- there are no “frontiers” of the usefulness of annotated corpora both for linguistic theory and NLP applications