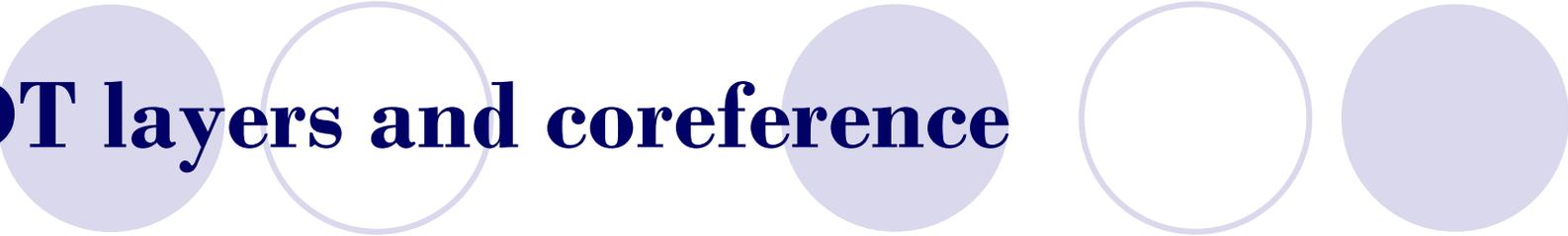




# **Coreferential Relations in the Prague Dependency Treebank**



**Eva Hajičová**



# 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

# Tectogrammatical annotation

- semiautomatic → user-friendly tree editor (TRED)
- 3 steps (phases):
  - build-up of underlying syntactic tree structures (incl. nodes deleted on the shallow structure) and assigning the nodes functional labels
  - adding the values of the topic-focus attribute
  - adding the coreferential links

# Annotation of coreference relations in PDT

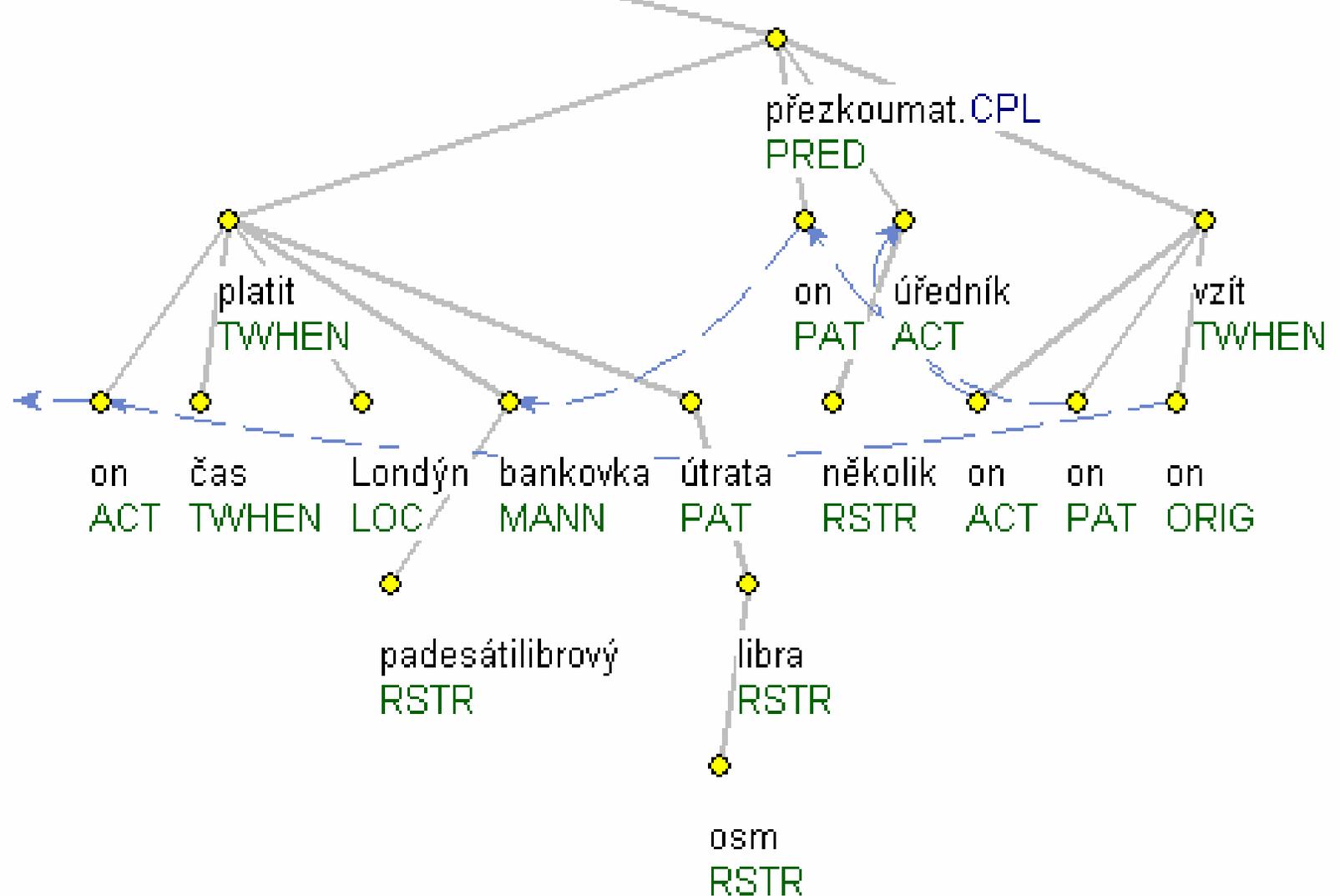
- coreference relations in the narrower sense
- a binary relation between an anaphor and an antecedent:
  - the antecedent may be in a different TGTS
  - the antecedent may also be an entity that is not represented in any TGTS
- 2 kinds of coreference
  - grammatical
  - textual

# Annotational scheme

- explicit coreference links are technically represented as pointers (pml 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

# 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



*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.*

# Grammatical coreference

- verbs (nouns, adjectives) of control
  - *John asked Mary to [0] come.*
- 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



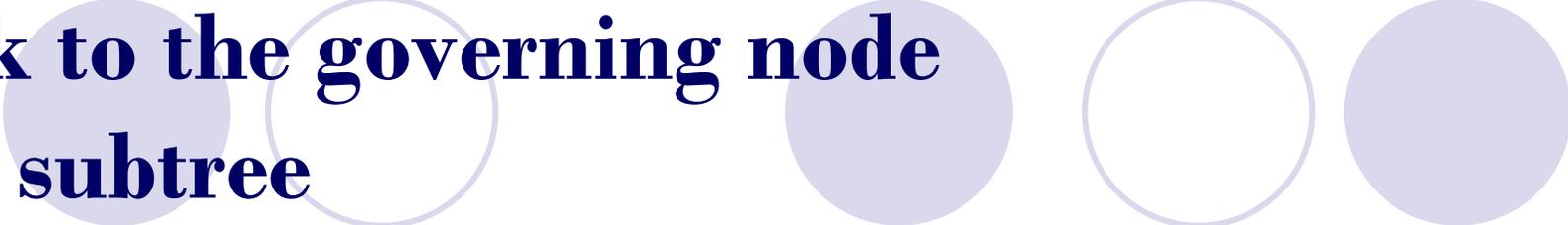
## Link to a particular node

- 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

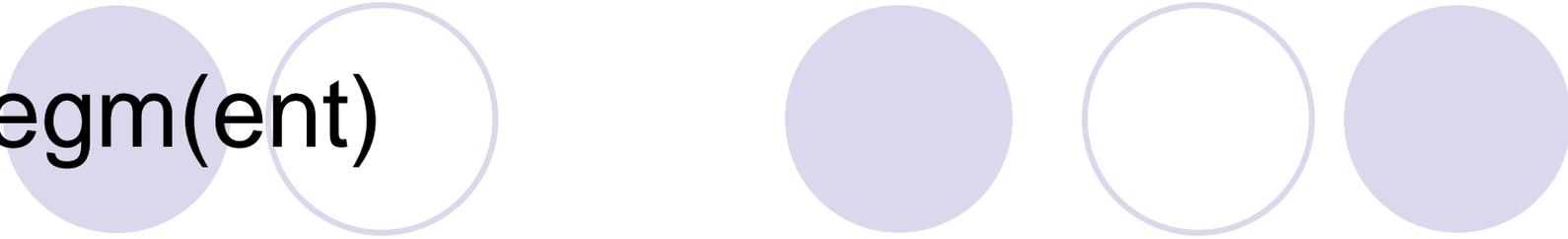


- antecedent is represented by this node plus (some of) its dependents; also the way how a link to a previous/following clause or a whole previous sentence is being established:

*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 30<sup>th</sup> this year. According to the information of LN, however, the radiator disappeared by the end of the last year.*

# 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 <b>textual</b> coreference	
1. #PersPron	18 622 (83%)
2. ten	3 733 (16.7%)
...	

# Statistics: expressed vs. restored

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

## 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.*

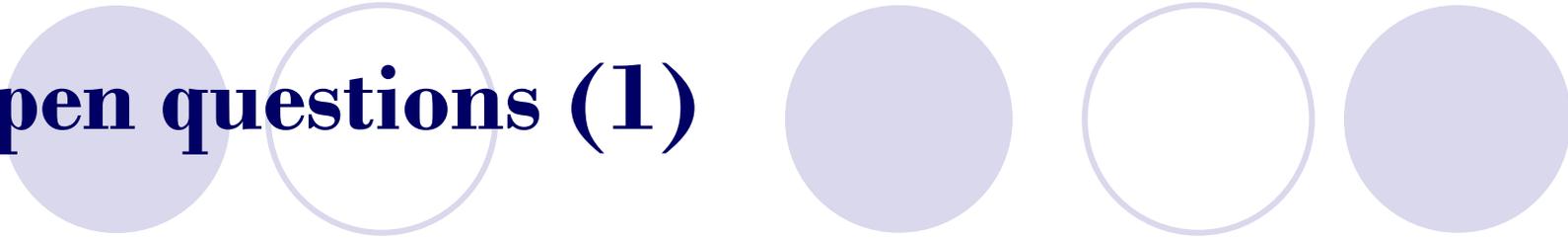
→ coreference to a segment („inferential“ type):

*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)

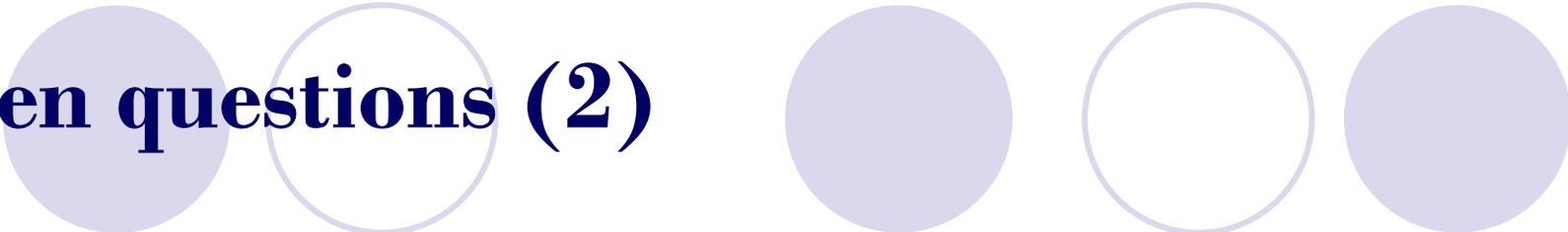


Coreferential link leads to the root

× 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)

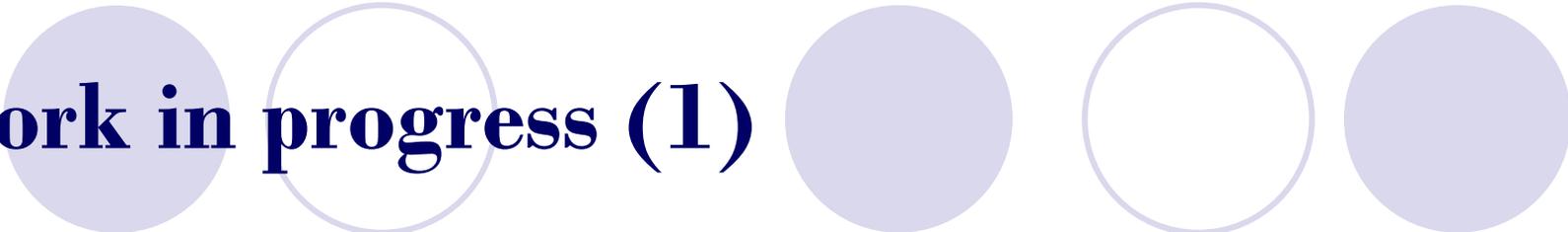
Nodes are reconstructed also with nominalizations:

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

*colouring* → Gen.ACT

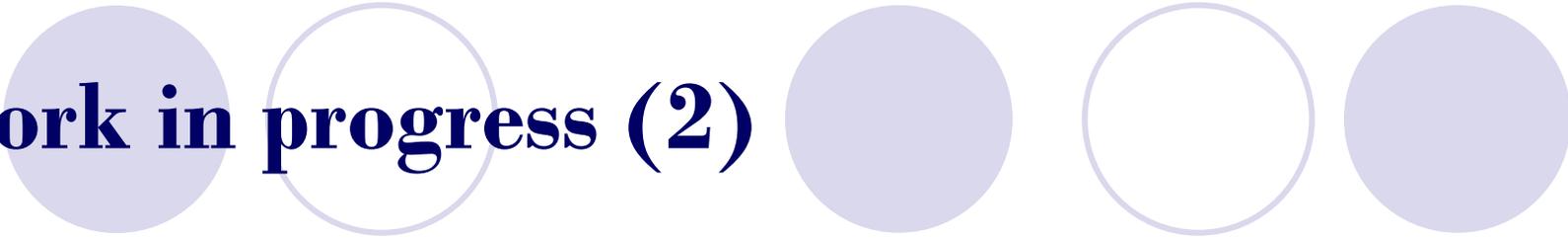
→ Gen.PAT → *on*.PAT → *slovo* [word]

# Work in progress (1)



- Nouns as anaphors: anaphoric relations leading from nouns
- Rough classification
  - Identity
  - Part and whole relation
  - Function
  - Other types (of bridging anaphora)

# Work in progress (2)



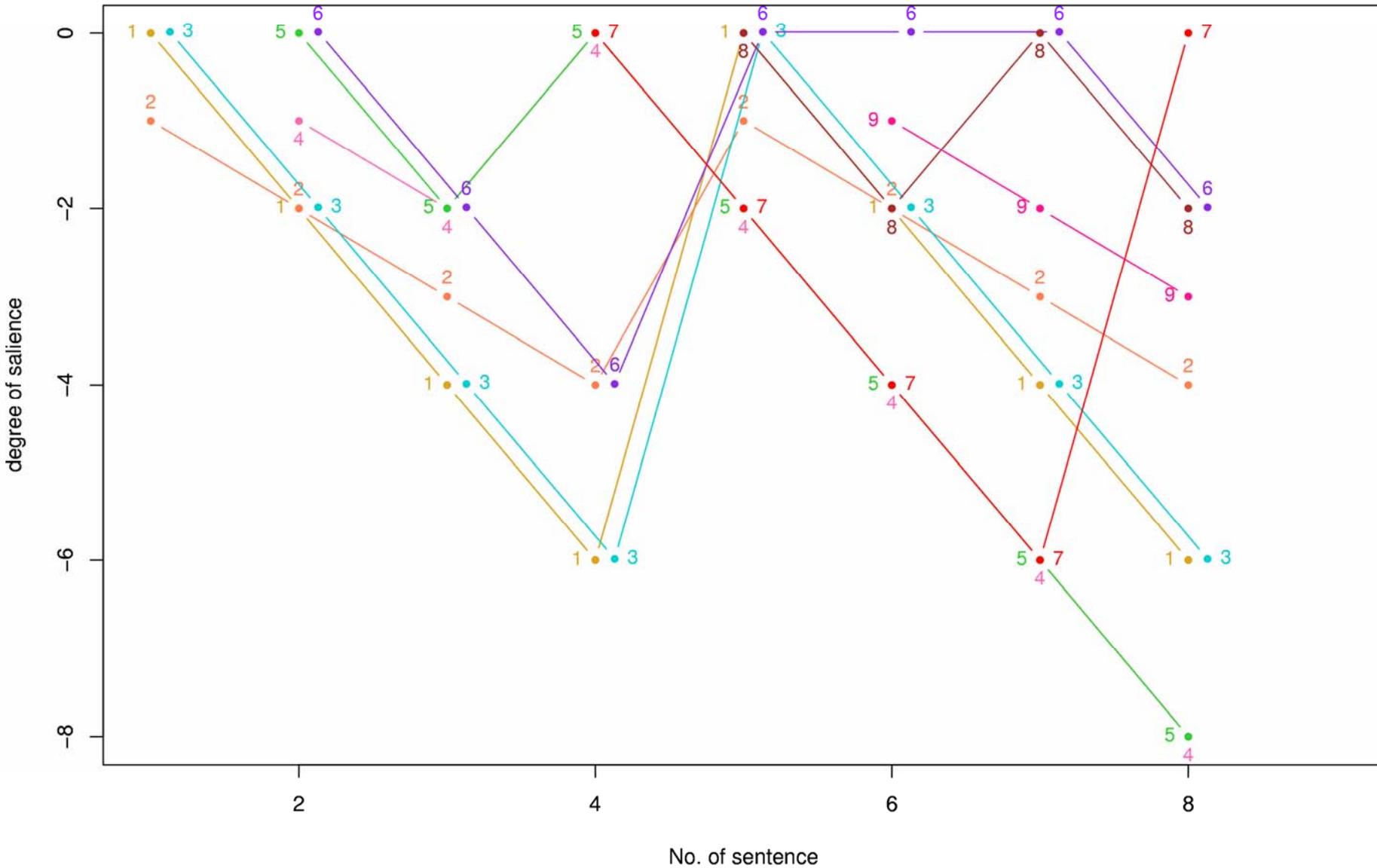
- Discourse structure analysis:
  - Hypothesis: A finite mechanism exists that enables the addressee to identify the referents on the basis of a partial ordering of the elements in the stock of knowledge (information) shared by the speaker and the addressees (according to the speaker's assumption), based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



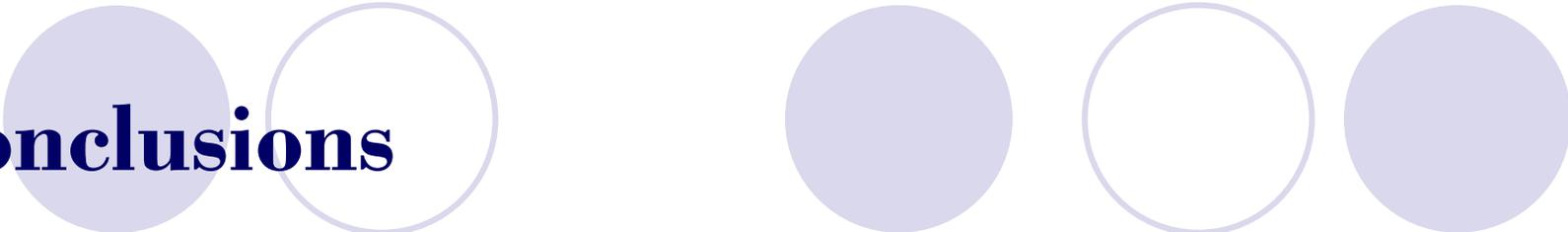
# Stock of shared knowledge

- SSH: a structured whole
- Hierarchy of activation of the SSK elements – a partial ordering
- Heuristic rules” for the assignment of degrees of activation based on:
  - TFA value
  - coreferential links
  - outer form (pronoun, full noun group)
- Implementation of the rules and visualization of the results

In95048\_092\_TF.dat



# 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