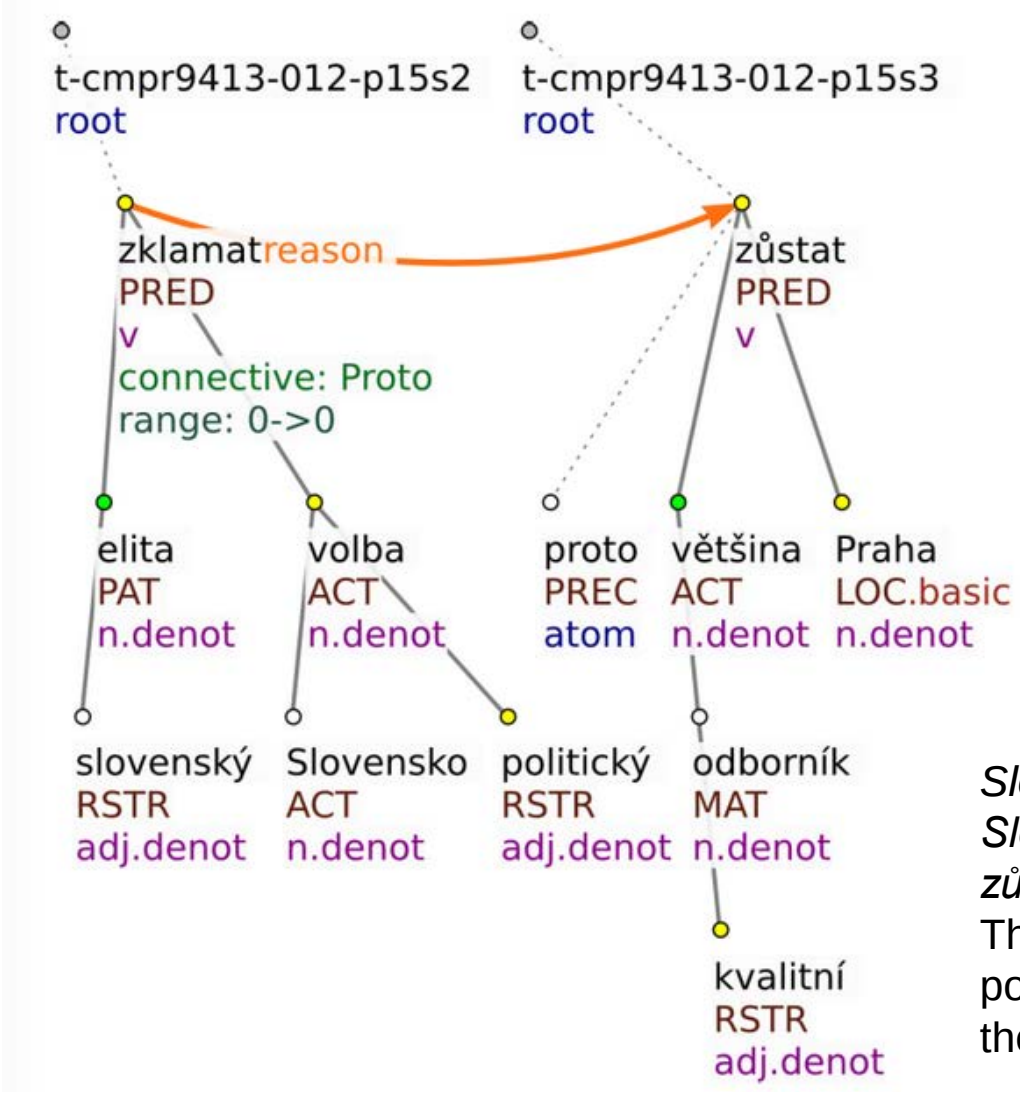
T-Layer of PDT: Discourse Relations and Other Annotation Added to PDT 3.0

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz

Discourse Annotation



Slovenská elita byla zklamána politickou volbou Slovenska. Proto většina kvalitních odborníků zůstala v Praze.
The Slovak elite were disappointed by the political choice of Slovakia. Therefore, most of the quality specialists stayed in Prague.

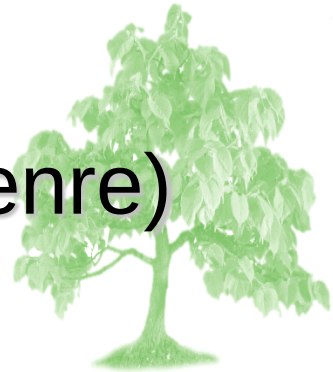
Discourse Annotation

- discourse connectives ("binary predicates")
  - inflexible
  - never act as "sentence members"
  - conjunctions – coord. (*and, but*) and some subord. (*because, if, while*), some particles (*also, only*), sentence adverbials (e.g. *afterwards*), …
  - must be explicitly expressed
- discourse units ("arguments") … finite clauses (basically)
- semantic relation of these units
  - 23 values: condition, gradation, opposition, purpose, reason, ….

- list structures
  "itemize", "enumerate"
- text phenomena –
  like article headings, captions, non-coherent texts (collection of news)

- structured attribute *discourse* (start node of the relation)
- (additional annotation is captured in *discourse_groups* and *discourse_special*)

Genre Specification (within journalistic genre)



- monological genres
  - critical review
  - letters from readers
  - cultural program
  - sport news
  - comment
  - news report
  - essay
  - weather forecast
  - …
- dialogue
  - topical interview
  - interview with a personality
- other
  - collection (various text in one document)
  - caption
  - metatext (text resulting from an error in corpus processing)
  - other (unclear, esp. for isolated sentences)

attribute attached to whole documents

Multiword Expressions (MWEs)

- info on MWES …
  - attribute *mwes* (root node of the tectogrammatical tree)
  - values: list of MWEs in the tree
  - for each MWE: *ID*, a *basic_form*, a *type* and a *list of identifiers of t-nodes*
  - 2 types of MWE
  - multiword lexeme (phraseme, light verb construction)
  - named entity

lexeme - a multiword lexeme
person - a name of a person or an animal
institution - an institution name
location - a geographical location
object - names of books, units of measurement, biological names of plants and animals
address - an address
time - date and time expressions
biblio - a bibliographic entry
foreign - a foreign expression
number - a numerical value, usually a range

Multiword Expressions (MWEs)

- info on MWES …
  - attribute *mwes* (root node of the tectogrammatical tree)
  - values: list of MWEs in the tree
  - for each MWE:
    ID, a *basic_form*, a *type* and a *list of identifiers of t-nodes*
  - 2 types of MWE
  - multiword lexeme (phraseme, light verb construction)
  - named entity

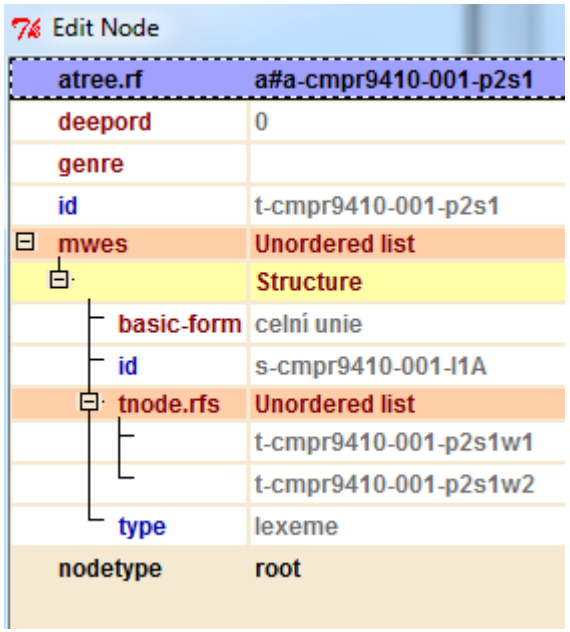
Prezident Havel by měl 15. července* na Pražském hradě** jmenovat třináct soudců Ústavního soudu***.

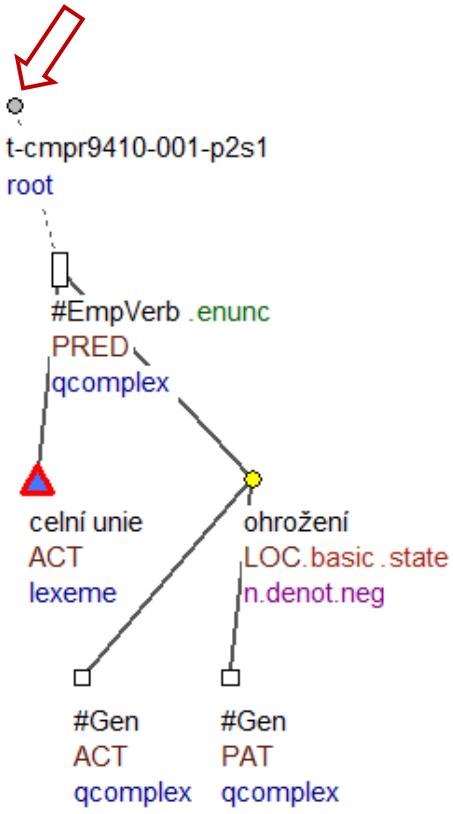
* – "15. July" – *date*, *basic_form* "15. červenec" (nominative case)
** – "at Prague Castle" (locative case) – *location*, *basic_form* "Pražský hrad" (nominative)
*** – "[of] Constitutional Court" (genitive) – *institution*, *basic_form* "Ústavní soud" (nominative)

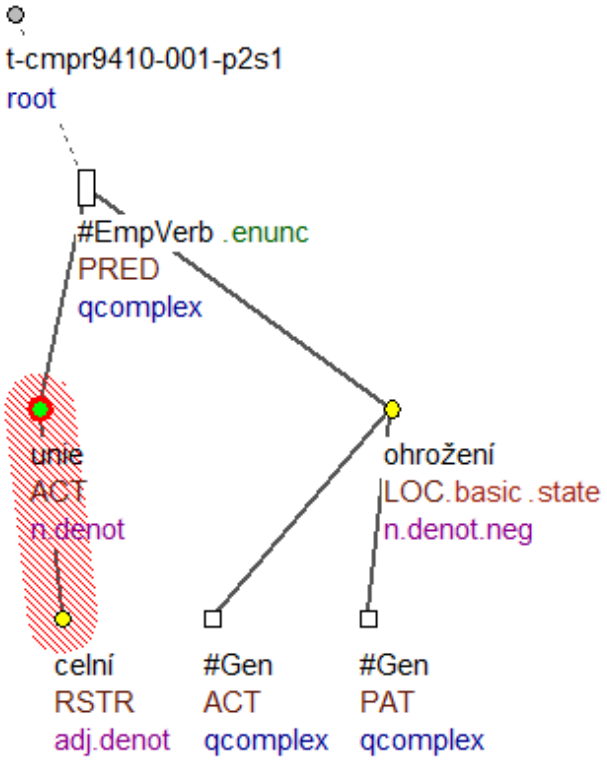
Funkce ústavního soudce* je neslučitelná s členstvím v politických stranách**.

* – "[of] constitutional judge" (genitive) – *lexeme*, *basic_form*"ústavní soudce" (nominative)
** – "in political parties" (instrumental, plural) – *lexeme*, *basic_form* "politická strana" (nominative, singular)

Multiword Expressions (MWEs)







Celní unie v ohrožení.
The Customs Union in danger.

References

- Manual for Tectogrammatical Annotation
  http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html
- Hajičová et al. (2002). *Úvod do teoretické a počítačové lingvistiky. I. sv. - Teoretická lingvistika*. Karolinum, Praha.
- Kučová, L., Hajičová, E., Veselá, K. Havelka, J. (2005) Topic-focus articulation and anaphoric relations: A corpus based probe. In *PBML* 84, pp. 5-12.
- Hajičová, E. (1999) The Prague Dependency Treebank: Crossing the Sentence Boundary". In *Proceedings of TSD'99*, Springer-Verlag Berlin Heidelberg, pp. 20-27