Prague Dependency Treebank:
Introduction – trees, dependency

Markéta Lopatková, Jiří Mírovský

Institute of Formal and Applied Linguistics, MFF UK
lopatkova@ufal.mff.cuni.cz

NPFL075 Prague Dependency Treebank

Lectures:

Markéta Lopatková Fri, S6, 14:00-15:30

Practical sessions:

Jiří Mírovský Fri, SU1, 12:20-13:50

http://ufal.mff.cuni.cz/course/npfl075

Requirements:

- Homework (40%)
- Activity (10%)
- Final test (50%)

Assessment:

- excellent (= 1) ≥ 90%
- very good (= 2) ≥ 70%
- good (= 3) ≥ 50%

Prague Dependency Treebank

Collection of:

- linguistically annotated data (Czech)
- tools and data format(s)
- documentation

Another point of view:

- annotation scheme
- framework for annotation of different languages
- underlying linguistic theory (Functional Generative Description)

Prague Dependency Treebank

Collection of:

- linguistically annotated data (Czech)
- tools and data format(s)
- documentation

Another point of view:

- annotation scheme
- framework for annotation of different languages
- underlying linguistic theory (Functional Generative Description)

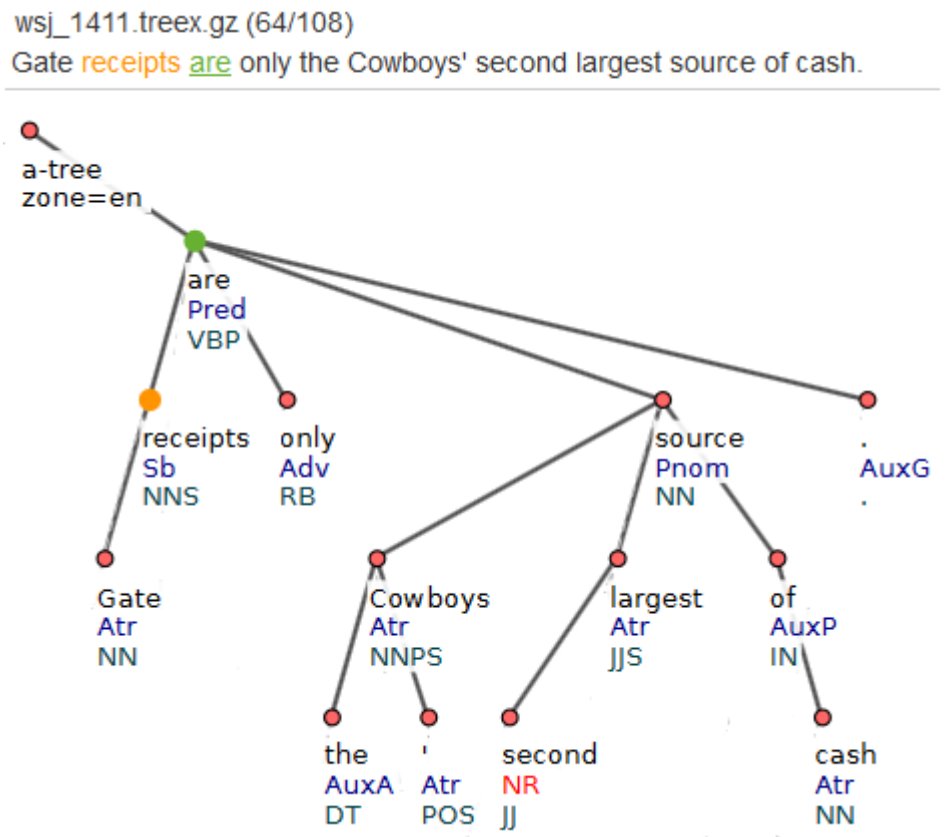
What about other/similar approaches:

- HamleDT
- Universal Dependencies

Outline of the lecture

- trees (graph theory and data format)
- phrase structure trees and dependency trees
- dependency and non-dependency relations
- non-projectivity

How to capture sentence structure?



Graph theory: tree

tree (graph theory):

definiftion:

- finite graph ⟨N, E⟩, N ~ nodes/vertices, E ~ edges $\{n_1,n_2\}$
- connected
- no cycles, no loops
- no more than 1 edge between any two different nodes

⇔ (undirected) graph

any two nodes are connected by exactly one simple path

Graph theory: tree

tree (graph theory):

definiftion:

- finite graph ⟨N, E⟩, N ~ nodes/vertices, E ~ edges $\{n_1,n_2\}$
- connected
- no cycles, no loops
- no more than 1 edge between any two different nodes

⇔ (undirected) graph

any two nodes are connected by exactly one simple path

rooted tree

- rooted ⇒ orientation (i.e., edges ordered pairs $[n_1,n_2]$)

Graph theory: tree

tree (graph theory):

definiftion:

- finite graph ⟨N, E⟩, N ~ nodes/vertices, E ~ edges $\{n_1,n_2\}$
- connected
- no cycles, no loops
- no more than 1 edge between any two different nodes

⇔ (undirected) graph

any two nodes are connected by exactly one simple path

rooted tree

- rooted ⇒ orientation (i.e., edges ordered pairs $[n_1,n_2]$)

directed tree … directed graph

- which would be tree
  - if the directions on the edges were ignored, or
  - all edges are directed towards a particular node ~ the ***root***

Data structure: tree

tree as a data structure:

- rooted tree (as in graph theory)
- all edges are directed from a particular node ~ the ***root***

Data structure: tree

tree as a data structure:

- rooted tree (as in graph theory)
- all edges are directed from a particular node ~ the ***root***

+

- (linear) ordering of nodes:
  the children of each node have a specific order

Data structure: tree (properties)

tree as a data structure:

- "tree-ordering" D … partial ordering on nodes
  $u \leq v \Leftrightarrow_{def}$ the unique path from the root to v passes through u
  (weak ordering ~ reflexive, antisymmetric, transitive)

- "linear ordering" … (partial) ordering on nodes
  (strong ordering ~ antireflexive, asymmetric, transitive)

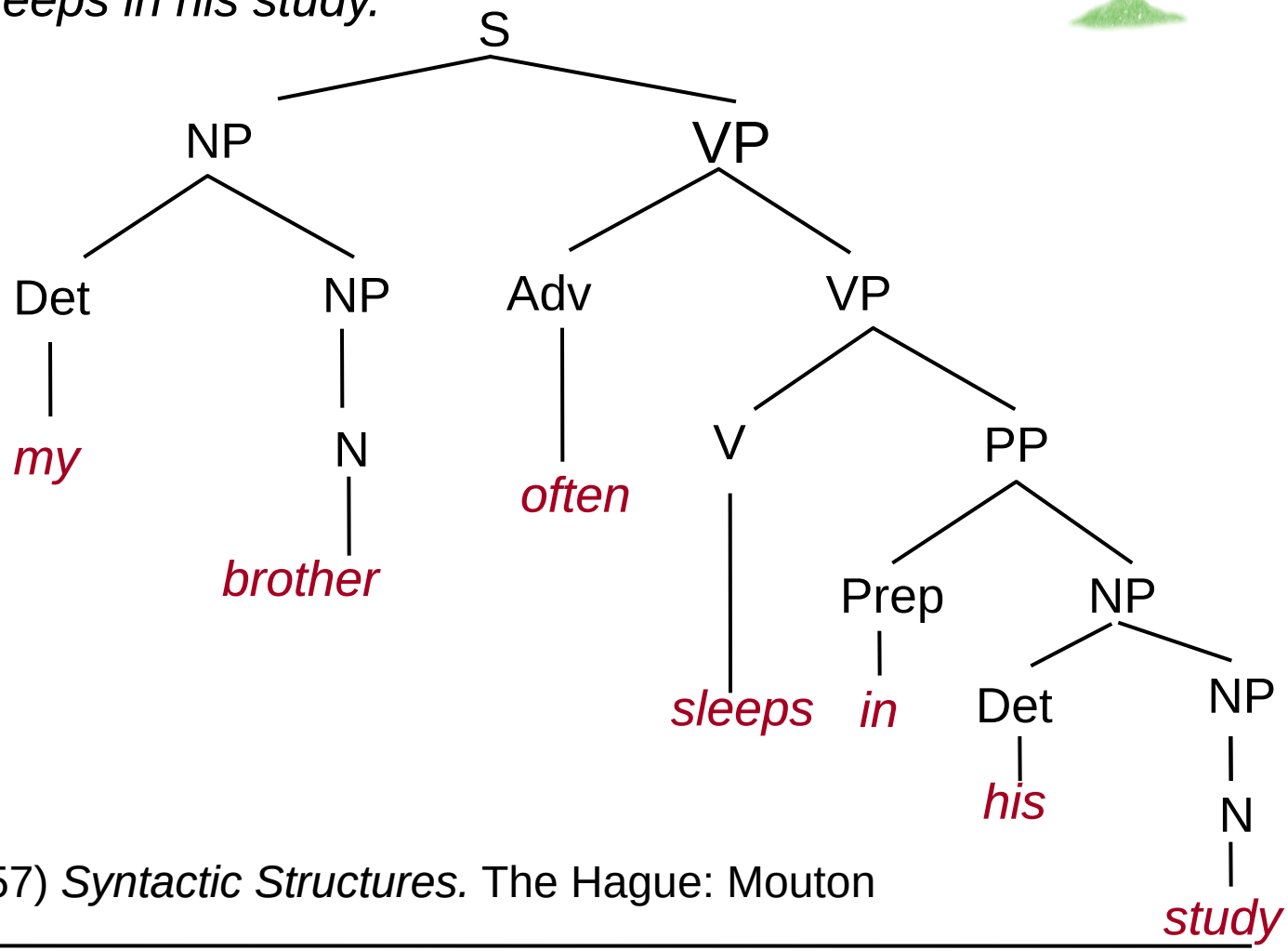
Tree-based structures in CL

two types of tree-based structures in CL:

- phrase structure tree / constituent structure tree
- dependency tree

Phrase structure tree

My brother often sleeps in his study.



Noam Chomsky (1957) *Syntactic Structures.* The Hague: Mouton

Phrase structure tree (definition)

$T = \langle N, D, Q, P, L \rangle$

$\langle N, D \rangle$ … ***rooted tree***

Q ... lexical and grammatical categories

L … labeling function $N \rightarrow Q$

D … oriented edges (branches)

~ relation on lex. and gram. categories

dominance relation

+

P ... relation on N ~ (partial strong linear ordering)

relation of *precedence*

Phrase structure tree (definition)

T = ⟨ N, D, Q, P, L ⟩

⟨N, D⟩ … ***rooted tree, directed***

Q ... lexical and grammatical categories

L … labeling function N → Q

D … oriented edges (branches)

~ relation on lex. and gram. categories

dominance relation

+

P ... relation on N ~ (partial strong linear ordering)

relation of *precedence*

+

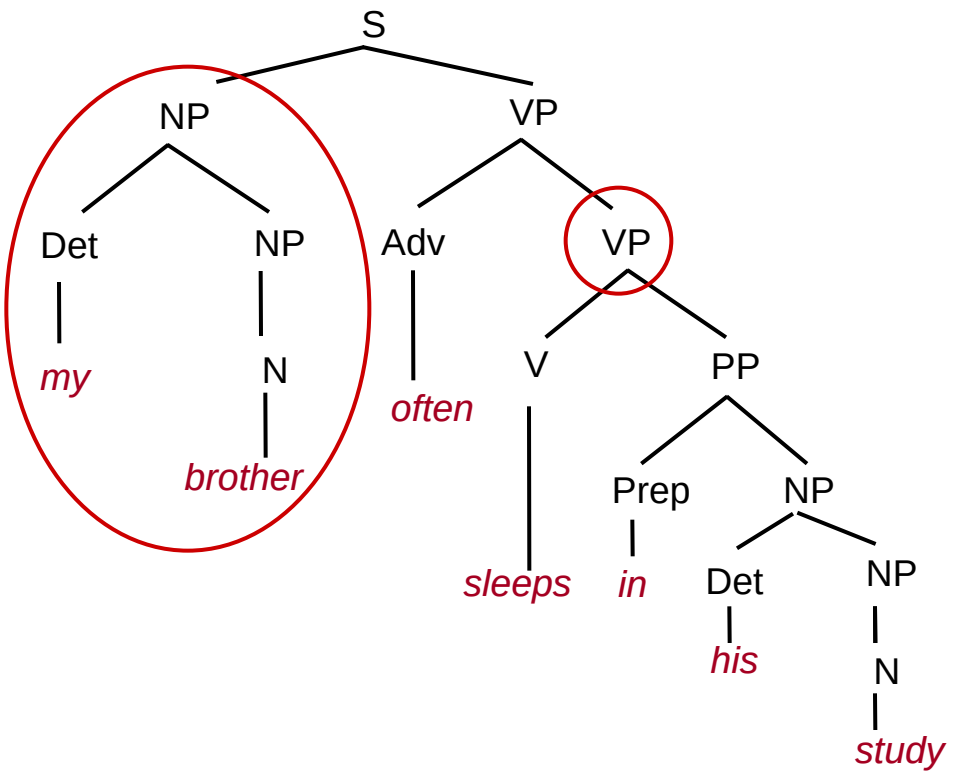
Relating dominance and precedence relations:

- *exclusivity* condition for D and P relations
- *'nontangling'* condition

Phrase structure tree (relation P)

- *exclusivity* condition for D and P relations

  $\forall\ x,y \in N$ holds: $([x,y] \in \mathrm{P} \vee [y,x] \in \mathrm{P}) \Leftrightarrow ([x,y] \notin \mathrm{D}\ \&\ [y,x] \notin \mathrm{D})$



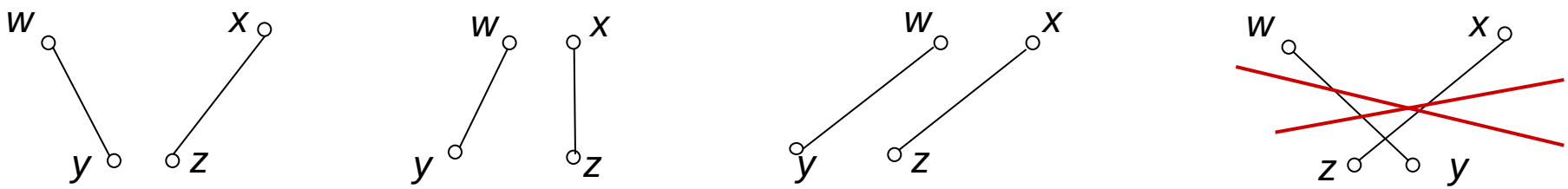
Phrase structure tree (relation P)

- *exclusivity* condition for D and P relations

  $\forall\ x,y \in N$ holds: $([x,y] \in \mathrm{P} \vee [y,x] \in \mathrm{P}) \Leftrightarrow ([x,y] \notin \mathrm{D}\ \&\ [y,x] \notin \mathrm{D})$

- *'nontangling'* condition

  $\forall\ w,x,y,z \in N$ holds: $([w,x] \in \mathrm{P}\ \&\ [w,y] \in \mathrm{D}\ \&\ [x,z] \in \mathrm{D}) \Rightarrow ([y,z] \in \mathrm{P})$



Phrase structure tree (relation P)

- *exclusivity* condition for D and P relations

  $\forall\ x,y \in N$ holds: $([x,y] \in P \vee [y,x] \in P) \Leftrightarrow ([x,y] \notin D\ \&\ [y,x] \notin D)$

- *'nontangling'* condition

  $\forall\ w,x,y,z \in N$ holds: $([w,x] \in P\ \&\ [w,y] \in D\ \&\ [x,z] \in D) \Rightarrow ([y,z] \in P)$

$\Rightarrow$

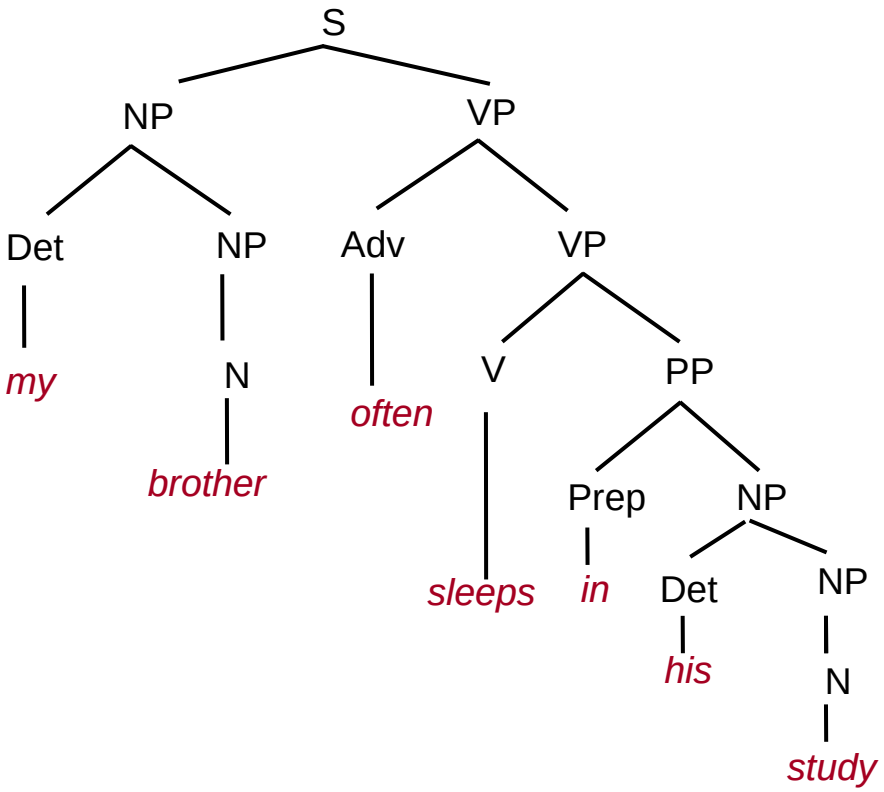
T = $\langle$ N,D,Q, P,L $\rangle$ phrase structure tree

- $\forall\ x,y \in N$ siblings $\Rightarrow [x,y] \in P$
- the set of its leaves is totally ordered by P

Phrase structure tree

Pros

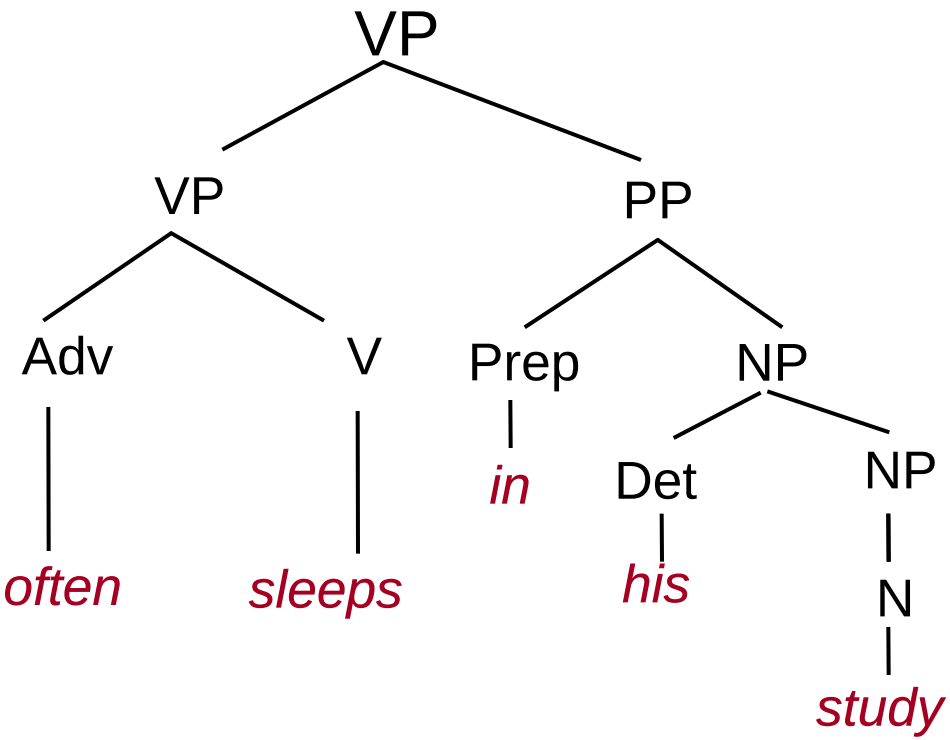
- derivation history / ‘closeness’ of a complementation
- coordination, apposition
- CFG-like
- derivation of a grammar



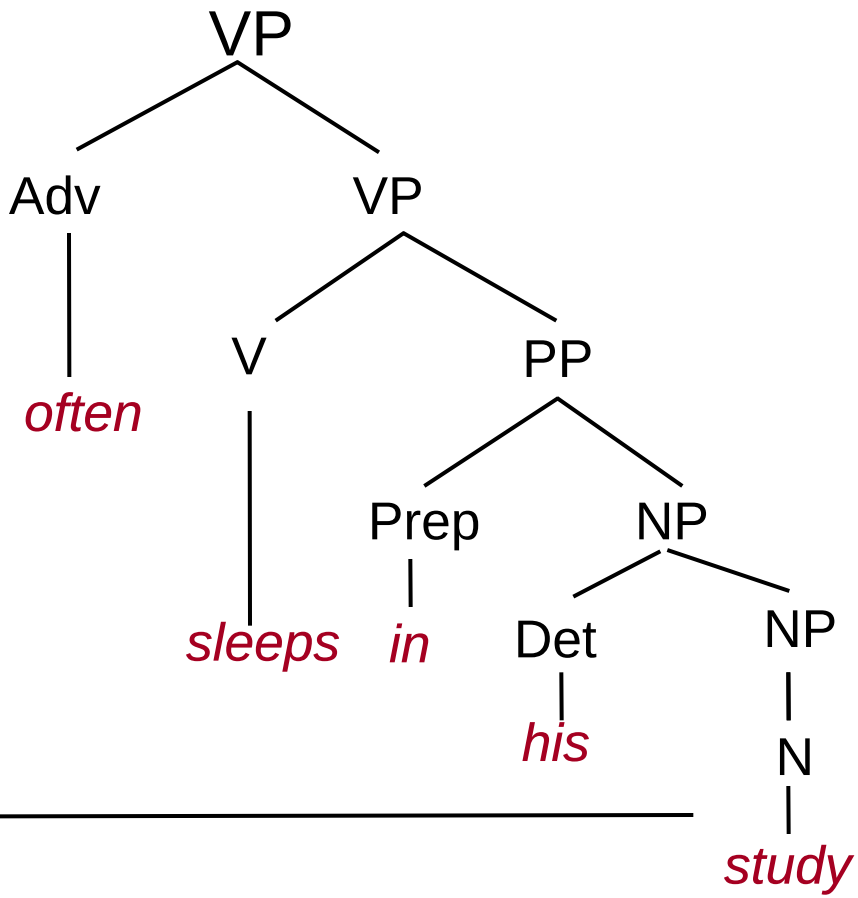
Phrase structure tree

derivation history / ‘closeness’:

… often sleeps in his study



… often sleeps in his study



Phrase structure tree

Pros

- derivation history / ‘closeness’ of a complementation
- coordination, apposition
- CFG-like
- derivation of a grammar

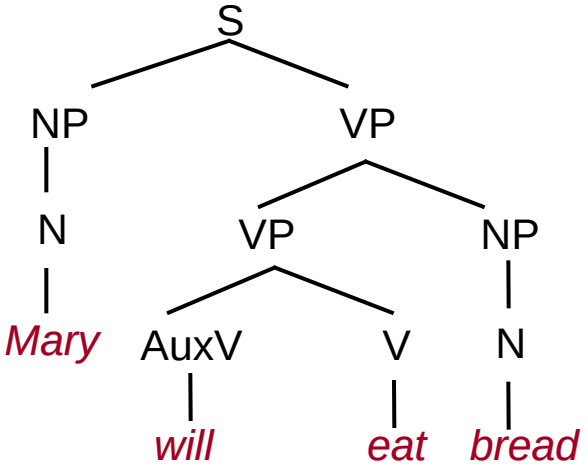
Contras

- complexity
  (number of non-terminal symbols)
- complement
  (‘two dependencies’)
  přiběhl bos
  [(he) arrived barefooted]
- *free word order*
  discontinuous ‘phrases’
  non-projectivity

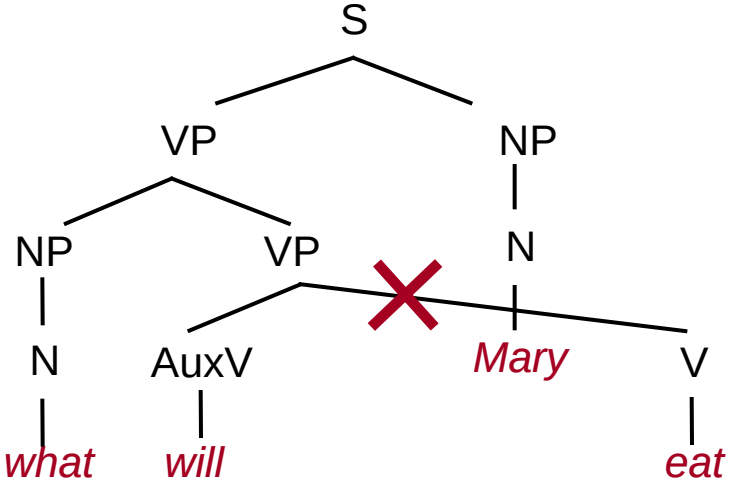
Phrase structure tree

discontinuous ‘phrases’: solution for English

Mary will eat bread.



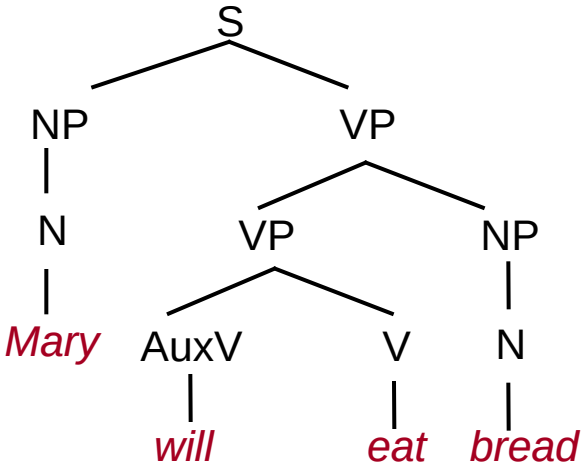
What will Mary eat?



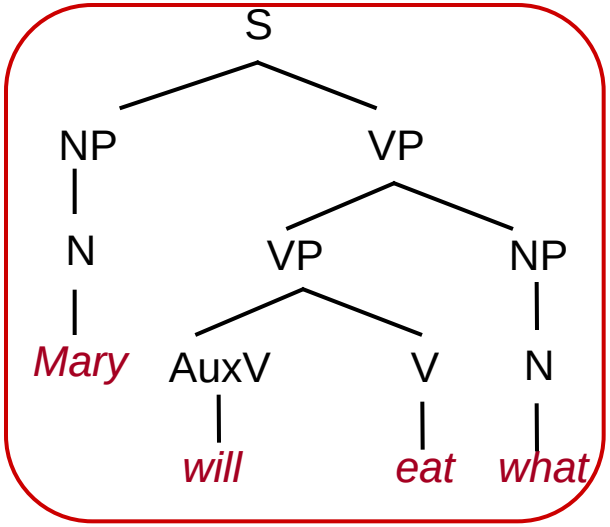
Phrase structure tree

discontinuous ‘phrases’: solution for English

Mary will eat bread.



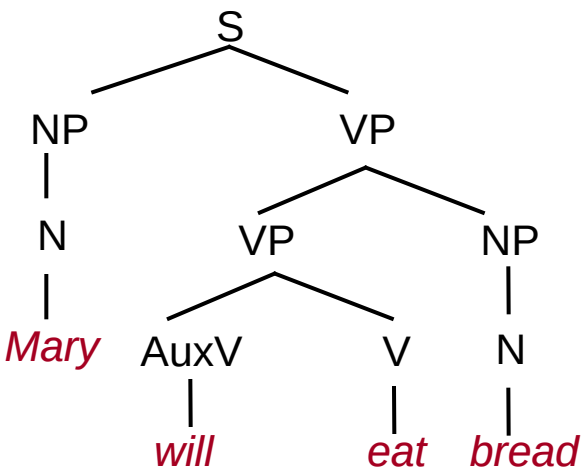
What will Mary eat?



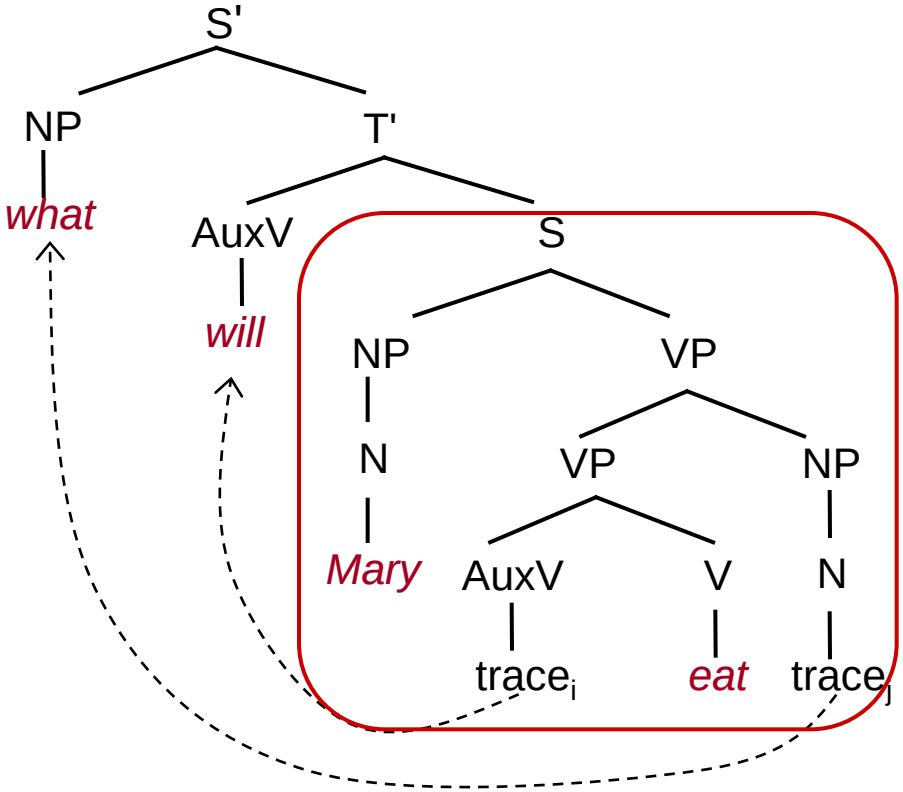
Phrase structure tree

discontinuous ‘phrases’: solution for English

Mary will eat bread.



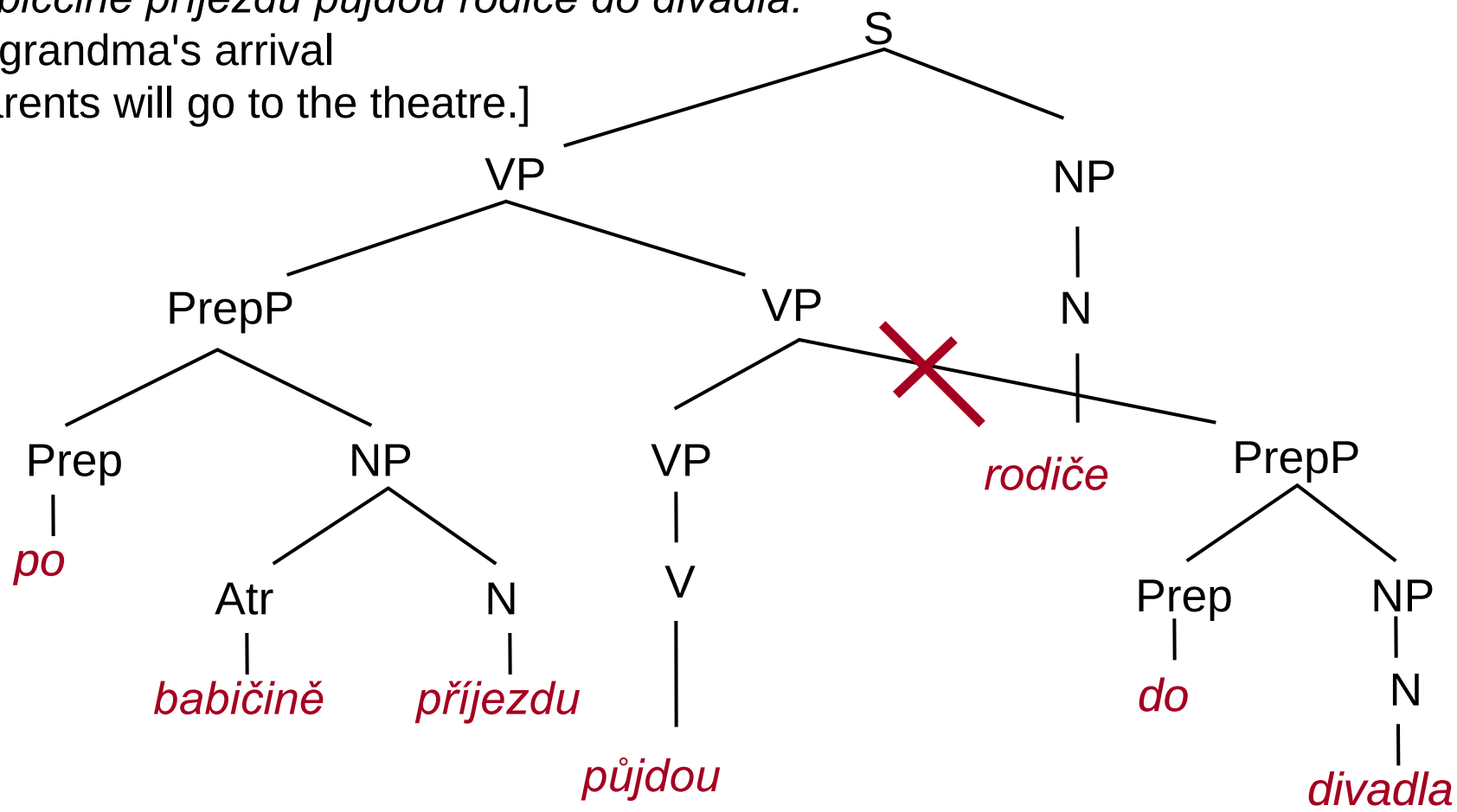
What will Mary eat?



Phrase structure tree

discontinuous ‘phrases’:

Po babiččině příjezdu půjdou rodiče do divadla.
[After grandma's arrival
the parents will go to the theatre.]



Corpora with phrase structure trees

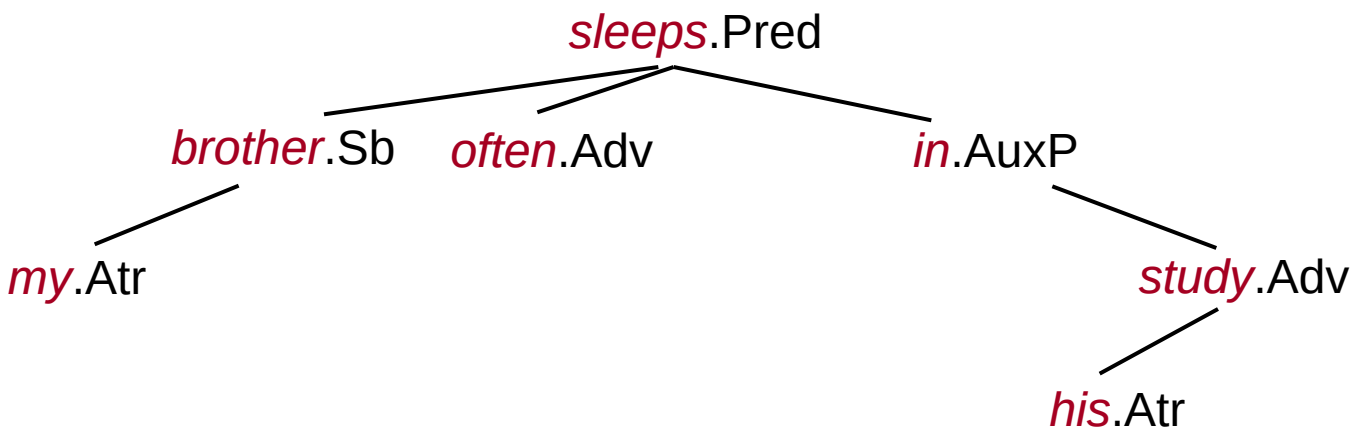
- Penn Treebank (1995)
  Mitchel Marcus (1993) Computational Linguistics, vol. 19
  http://www.cis.upenn.edu/~treebank/

  Penn Arabic Treebank, Penn Chinese Treebank
- International English Treebank (ICE)
  http://ice-corpora.net/ice/index.htm
- Paris 7
  http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php
- Szeged Treebank 2.0
  http://www.inf.u-szeged.hu/projectdirs/hlt/en/Szeged%20Treebank%202.0_en.html
- many many others

Dependency tree

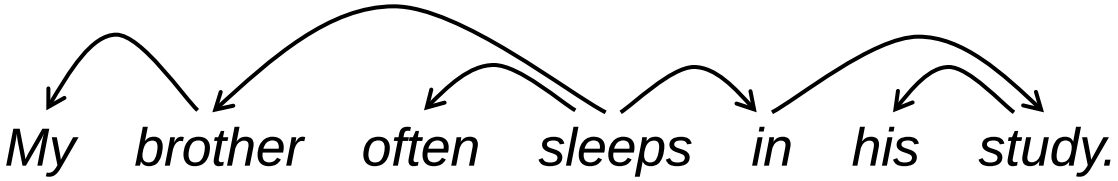
Dependency tree

My brother often sleeps in his study.



Lucien Tesnière (1959) *Éléments de syntaxe structurale.* Editions Klincksieck.

Igor Mel'čuk (1988) *Dependency Syntax: Theory and Practice.* State University of New York Press.



Dependency tree (definition)

T = ⟨ N, D, Q, WO, L ⟩

⟨N, D⟩ … ***rooted tree, directed***

Q ... lexical and grammatical categories

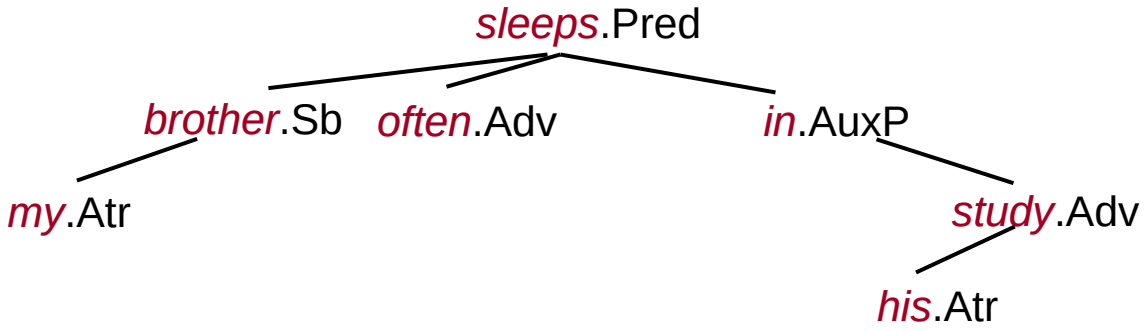
L … labeling function N → Q

D … oriented edges ~ relation on lex. and gram. categories

'dependency' relation

WO ...relation on N ~ (strong total ordering on N) …

word order



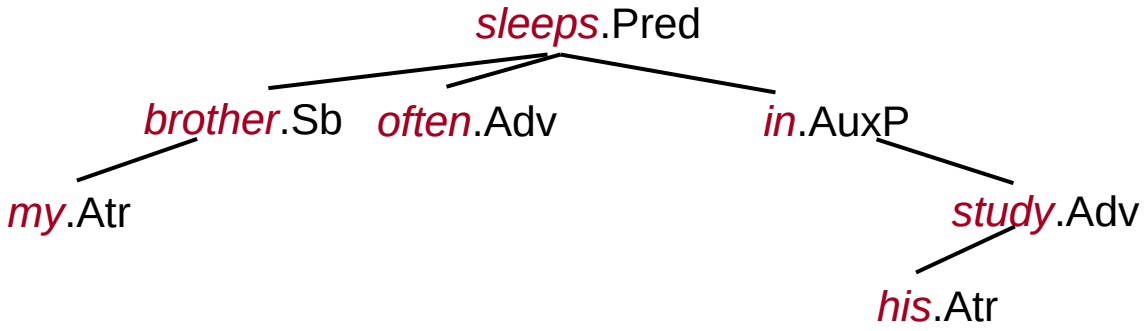
Dependency tree

Pros

- economical, clear
  (complex labels, ‘word’~ node)
- free word order
- head of a phrase

Contras

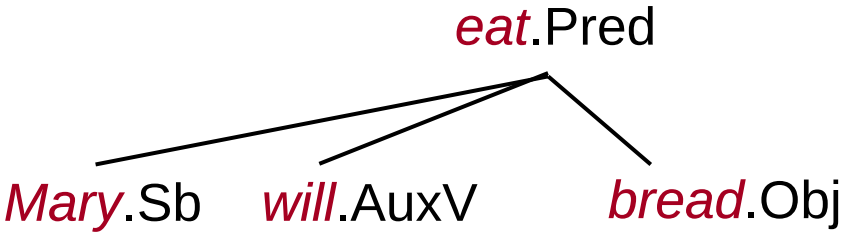
- no derivation history / 'closeness'
- coordination, apposition
- complement



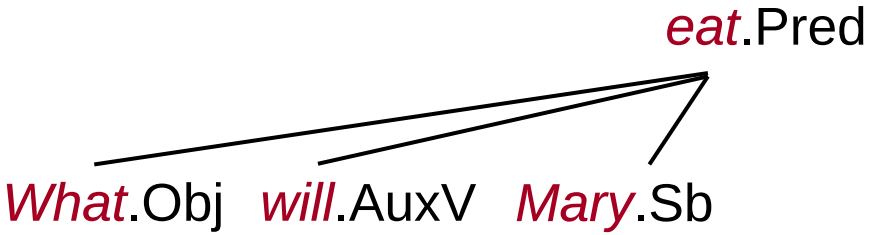
Dependency tree

discontinuous ‘phrases’: no problem

Mary will eat bread.

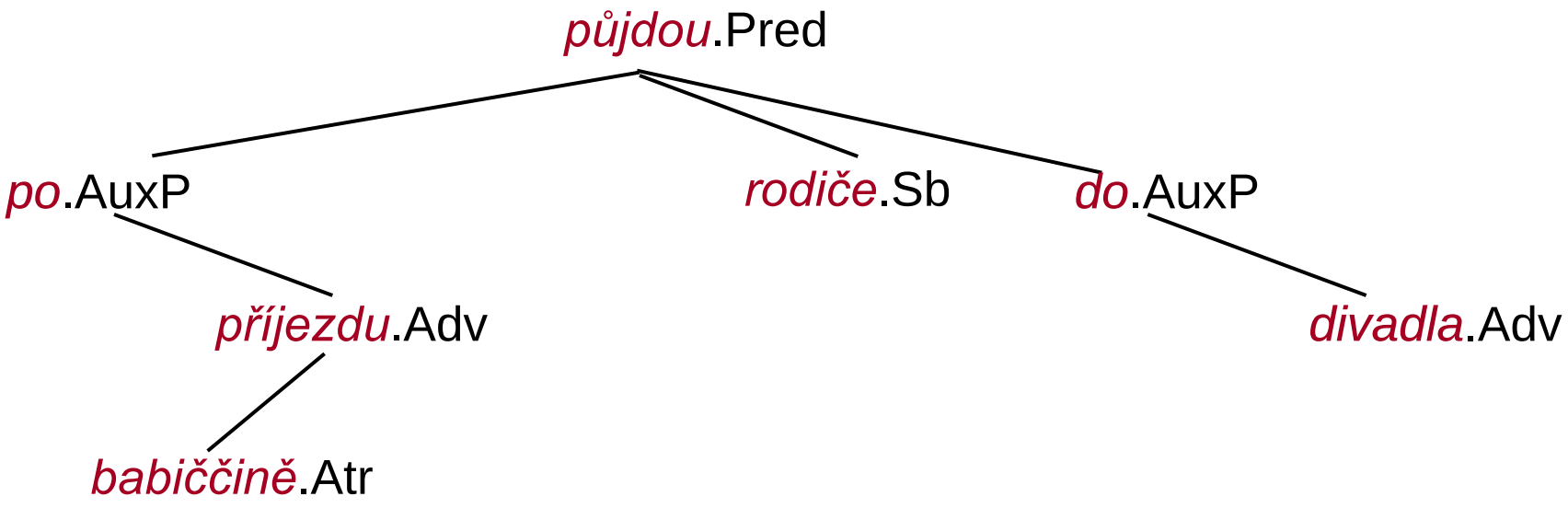


What will Mary eat?



Dependency tree

Po babiččině příjezdu půjdou rodiče do divadla.
[After grandma's arrival the parents will go to the theatre.]



Corpora with dependency trees

- PropBank (1995)
  http://propbank.github.io/
- family of Prague dependency treebanks: Czech, Arabic, English
  http://ufal.mff.cuni.cz/pdt.html
- HamleDT project (from 2012) http://ufal.mff.cuni.cz/hamledt
- Universal Dependencies http://universaldependencies.org/

- Danish Dep. Treebank
  http://mbkromann.github.io/copenhagen-dependency-treebank/
- Finnish: Turku Dependency Treebank
  http://bionlp.utu.fi/fintreebank.html
- Negra corpus
  http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html
- TIGERCorpus
  http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html/
- SynTagRus Dependency Treebank for Russian

Dependency and non-dependency relations

Dependency and non-dependency relations

edges ~ *dependency relations* (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- long discussion, number of linguistic criteria

  e.g., each complete subtree must be a "constituent", i.e., it must allow for several constructions like topicalization, proform substitution, ….;

*Mary **will eat** bread..*

⇨ lexical verb should be a dependent

Topicalization:
*… and **eat** Mary certainly will.*

Proform substitution:
Mary will do so. (do=eat)

Answer fragment:
What will Mary do? Eat.

VP-ellipsis:
Peter will eat and Mary will, too.

Dependency and non-dependency relations

edges ~ *dependency relations* (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- PDT criterion: *possible reduction*

  … dependent member of the pair may be deleted

  while the distributional properties are preserved (→ correctness is preserved)

Dependency and non-dependency relations

edges ~ *dependency relations* (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- PDT criterion: *possible reduction*

  … dependent member of the pair may be deleted

  while the distributional properties are preserved (→ correctness is preserved)

  - endocentric constructions … OK

malý stůl → stůl	*small table → table*
přišel včas → přišel	*he came in time → he came*
(přišel) velmi brzo → (přišel) brzo	*(he came) very soon → (he came) soon*

Dependency and non-dependency relations

edges ~ *dependency relations* (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- PDT criterion: *possible reduction*

  … dependent member of the pair may be deleted

  while the distributional properties are preserved (→ correctness is preserved)

  - endocentric constructions … OK
  - exocentric constructions … *principle of analogy* on word classes

Prší. [(It) rains.] … ∃ subjectless verbs
⇒ *Král zemřel.* [The king died.] … a verb rather than a noun is the head

The girl painted a bag. → *The girl painted.* ... ∃ objectless verbs
⇒ *The girl carried a bag* … an object is considered as depending on a verb

Dependency and non-dependency relations

edges ~ *dependency relations* (prototypically)

- dependency relation: binary relation
- governing/modified unit (head) – dependent/modifying unit (modifier)
- PDT criterion: *possible reduction*

  … dependent member of the pair may be deleted

  while the distributional properties are preserved (→ correctness is preserved)

  - endocentric constructions … OK
  - exocentric constructions … *principle of analogy* on word classes

  PLUS technical considerations

  e.g.: prepositions are below nouns;
  auxiliary verbs are (typically) below content verbs

Dependency and non-dependency relations

BUT also other relations:

coordination … "multiplication" of a single syntactic position

- different referents
- coordination of sentence members / sentences

  My sister Mary and John came late.
  Mary came in time but John was late.
  I can't leave since it hasn't stopped raining yet.
  Nemohu odejít, neboť ještě nepřestalo pršet.

- coordination may be embedded

  nice and romantic towers and castles
  krásné a romantické hrady a zámky

Dependency and non-dependency relations

BUT also other relations:

coordination … "multiplication" of a single syntactic position

- different referents
- coordination of sentence members / sentences

  My sister Mary and John came late.
  Mary came in time but John was late.
  I can't leave since it hasn't stopped raining yet.
  Nemohu odejít, neboť ještě nepřestalo pršet.

- coordination may be embedded

  nice and romantic towers and castles
  krásné a romantické hrady a zámky

apposition … "multiplication" of a single syntactic position

- identical referent

  Charles IV, Holy Roman Emperor
  The Hobbit, or There and Back Again

Dependency and non-dependency relations

BUT also other relations:

coordination … "multiplication" of a single syntactic position

- different referents
- coordination of sentence members / sentences
- coordination may be embedded

apposition … "multiplication" of a single syntactic position

- identical referent

⇨ necessary to enrich the data structure

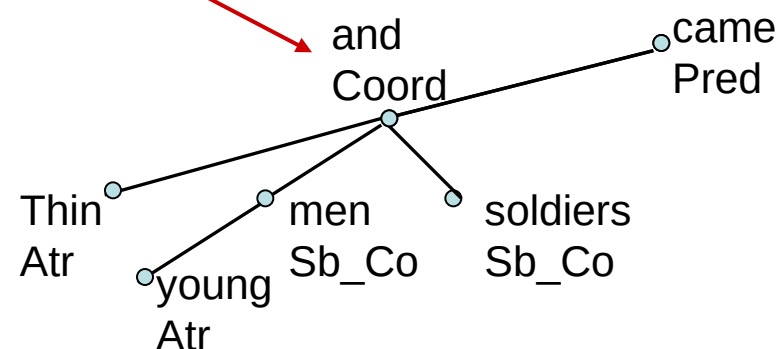
Coordination/apposition in dependency trees

PDT 2.0:

'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- *connecting node* (= node for coordinating / appositing conjunction)



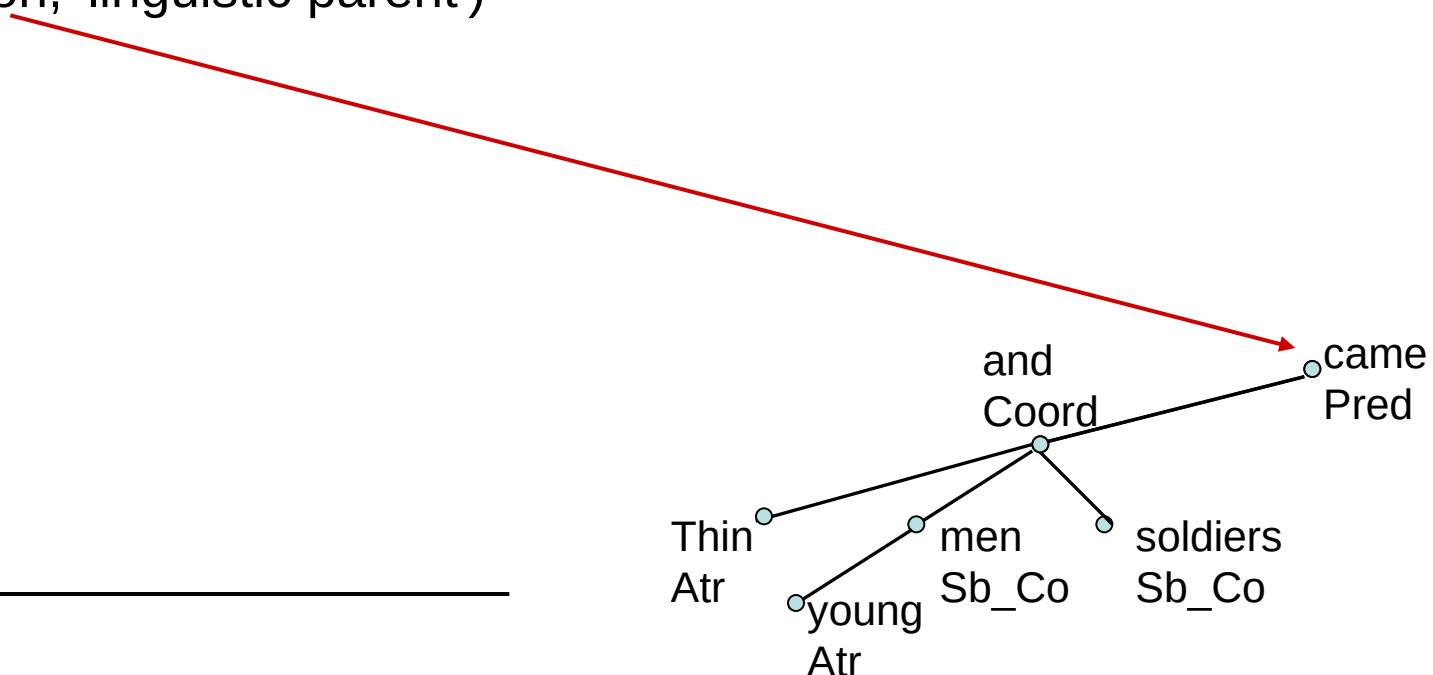
Coordination/apposition in dependency trees

PDT 2.0:

'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- *connecting node* (= node for coordinating / appositing conjunction)
- *effective parent* (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')



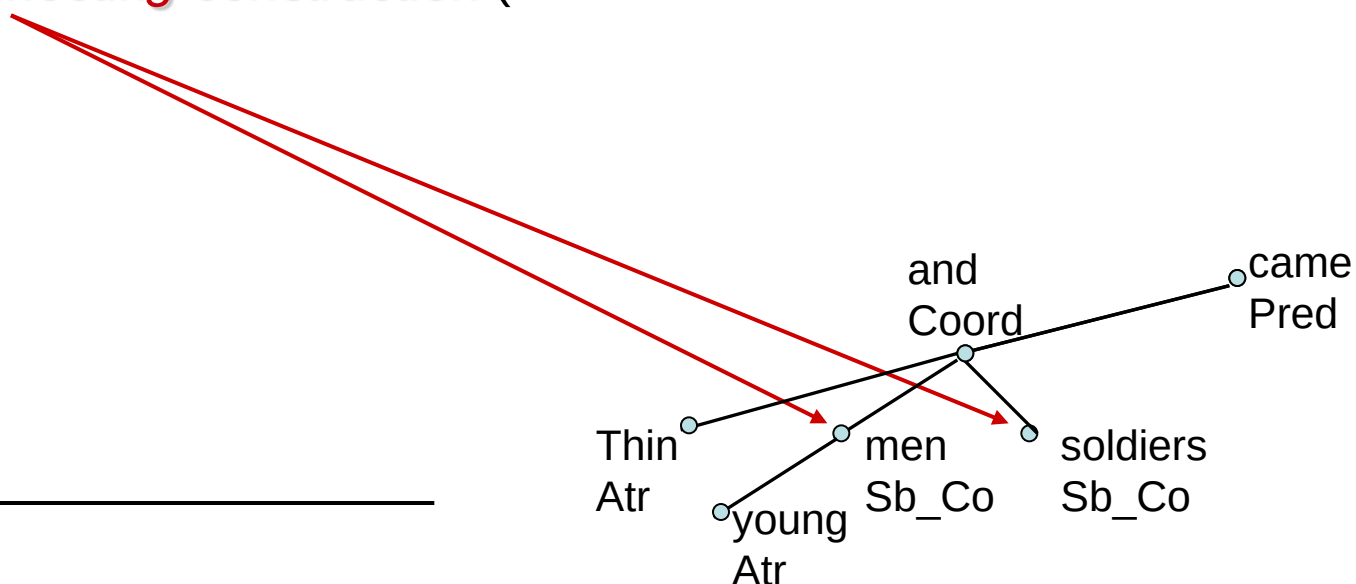
Coordination/apposition in dependency trees

PDT 2.0:

'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- *connecting node* (= node for coordinating / appositing conjunction)
- *effective parent* (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
- *members of a connecting construction* (= nodes that are coordinated / are in apposition)
    - `is_member`



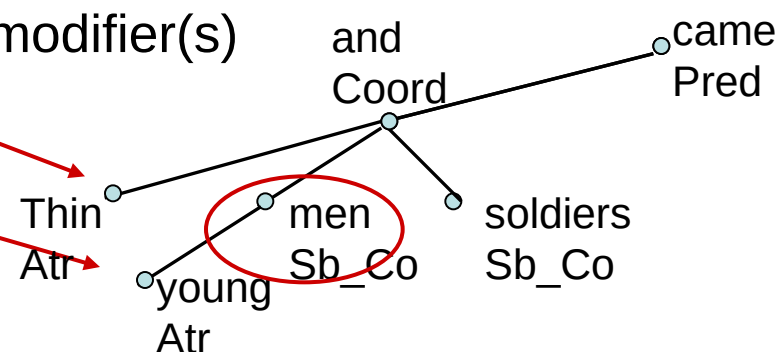
Coordination/apposition in dependency trees

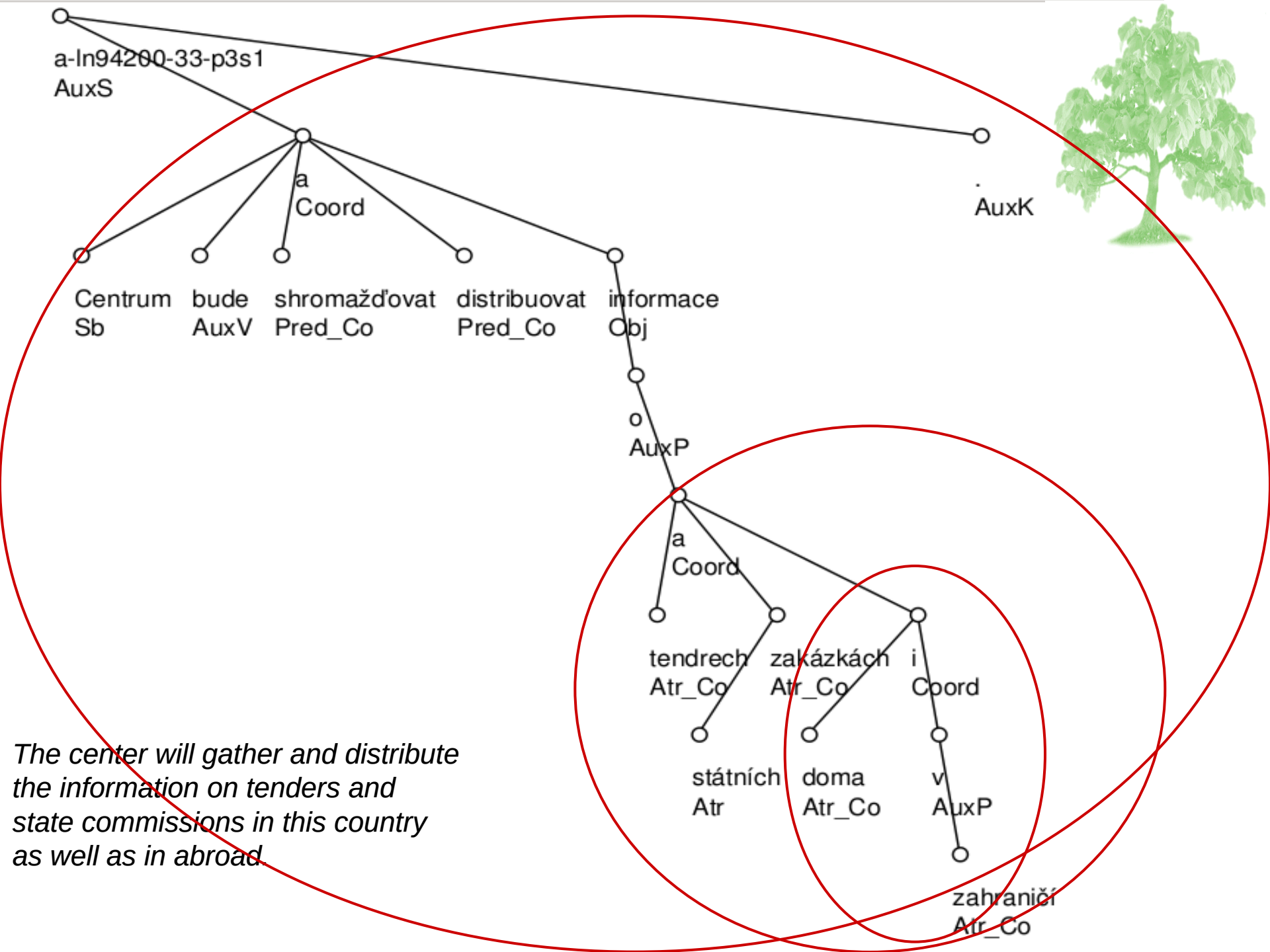
PDT 2.0:

'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- *connecting node* (= node for coordinating / appositing conjunction)
- *effective parent* (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
- *members of a connecting construction* (= nodes that are coordinated / are in apposition)
  - `is_member`
- *effective child(ren)* … modification(s) of the individual member of the connecting construction **+** common/shared modifier(s)
- *'pass-through' nodes*





The center will gather and distribute the information on tenders and state commissions in this country as well as in abroad.

Coordination/apposition in dependency trees

PDT 2.0:

- embedded connecting constructions ⇨ recursivity

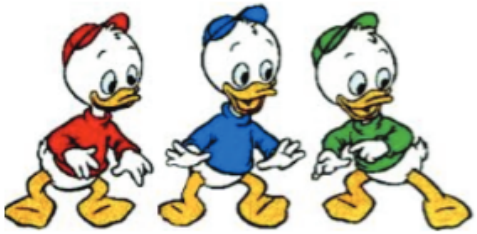
- *TrEd* (Tree Editor, Pajas):

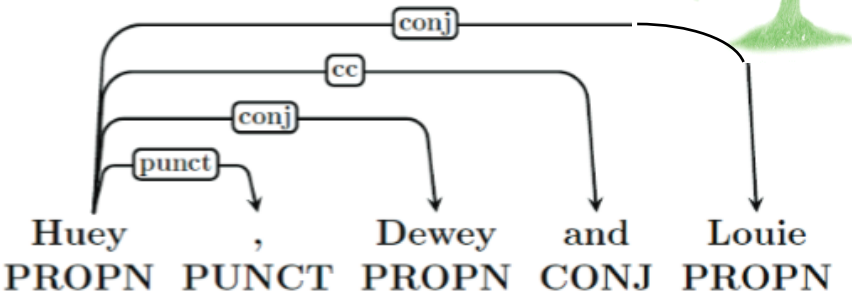
functions `GetEChildren, GetEParents`

Coordination/apposition in dependency

Universal Dependencies:

version 1 (2014):





(Slides stolen from Daniel Zeman)

the first conjunct

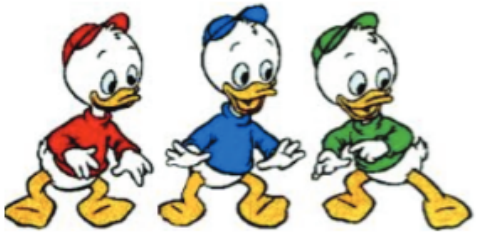
~ the head of all following conjuncts

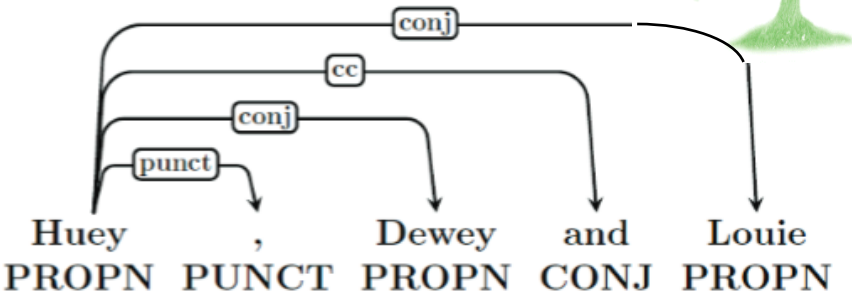
~ the head of any intervening coordinating conjunctions and punctuation

Coordination/apposition in dependency

Universal Dependencies:

version 1 (2014):

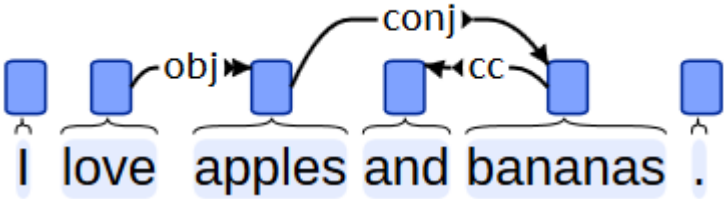




(Slides stolen from Daniel Zeman)

the first conjunct
~ the head of all following conjuncts
~ the head of any intervening coordinating conjunctions and punctuation

version 2 (2016):

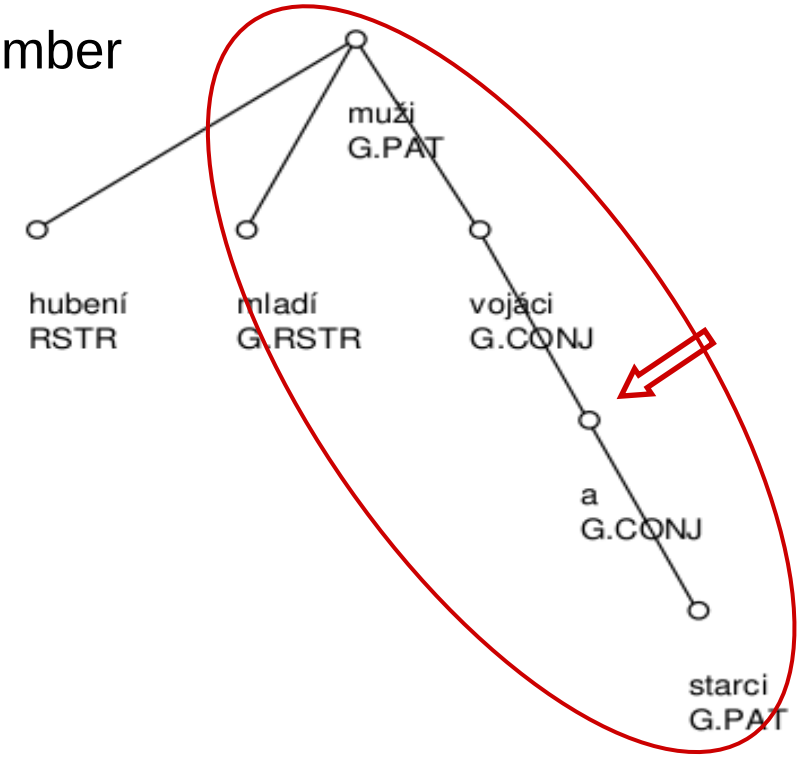


- the first conjunct ~ the head of all following conjuncts
- attach coordinating conjunctions and punctuation
  to the immediately succeding conjunct (instead of the first)

Coordination/apposition in dependency trees

Mel'čuk (1988):

- ‘grouping’ (G) … treating the first conjunct as the head
- problem: shared modification
  vs. modification of a single member

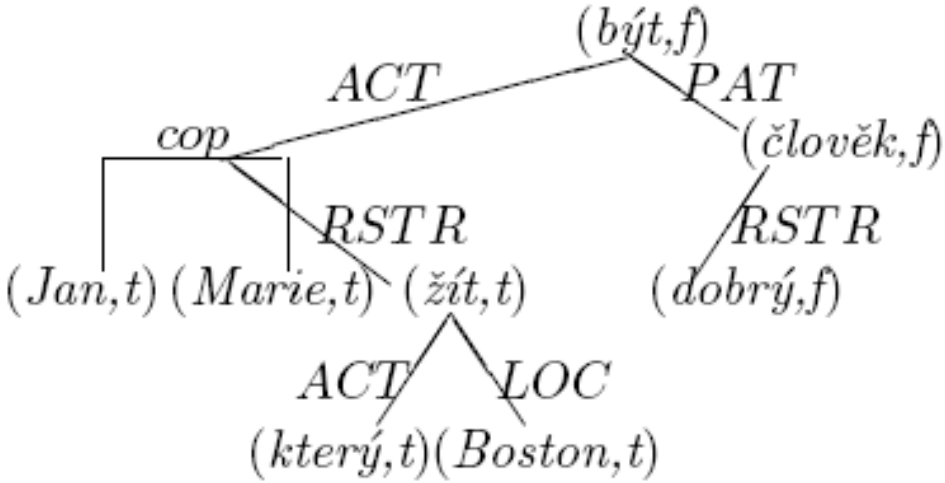


Hubení ((mladí muži) , vojáci a starci)
[Thin young men, soldiers and old-men]

Coordination/apposition in dependency trees

Petkevič (1995) … formal representation of FGD
two types of brackets for tree linearization:

- ⟨ ⟩ for dependencies
- [ ] for coordination



$$\langle[(Jan,t);\ (Marie,t)]_{cop}\ {}_{RSTR}\langle\langle(který,t)\rangle_{ACT}\ (žít,t)\ {}_{LOC}\langle(Boston,t)\rangle\rangle\rangle_{ACT}\ (být,f)\ {}_{PAT}\langle\langle(dobrý,f)\rangle_{RSTR}\ (člověk,f)\rangle$$

References

- Hajičová, E., Havelka, J., Sgall, P., Veselá, K., Zeman, D. (2004) Issues of Projectivity in the Prague Dependency Treebank. *PBML,* vol. 81
- Holan, T., Kuboň, V., Oliva, K., Plátek, M. (2000) On Complexity of Word Order. *Les grammaires de dépendance – Traitement automatique des langues,* vol. 41, no. 1, 273-300
- Kuhlmann, M., Nivre, J. (2006) Mildly Non-Projective Dependency Structures. In COLING/ACL Main Conference Poster Sessions, 507–514.
- Mel'čuk, I. (1988) *Dependency Syntax: Theory and Practice.* State University of New York Press, Albany
- Partee, B. H.; ter Meulen, A.; Wall, R. E. (1990) *Mathematical Methods in Linguistics*. Kluwer Academic Publishers
- Petkevič, V. (1995) A New Formal Specification of Underlying Structure. *Theoretical Linguistics*, vol. 21, No.1
- Sgall, P., Hajičová, E., Panevová, J. (1986) *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht/Academia, Prague
- Štěpánek, J. (2006) *Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu*. PhD Thesis, MFF UK

Dependency and non-dependency relations

other non-dependency relations in PDT:

- technical root – effective root of a sentence
- syntactically unclear expressions

  rhematizers; sentence, linking and modal adverbial expressions, conjunction modifiers
- list structures

  names, foreign expressions
- phrasemes