Zachycení (nejen) koordinací v závislostních stromech

Markéta Lopatková

ÚFAL MFF UK

Seminář Rozpoznávání a syntaktická analýza

listopad 2015

Natural language syntax: Treebanks

- text corpora, esp. treebanks
  - tens of languages
  - stress on morphology, syntax
  - manual or (semi)automatic annotation

⇨ millions of words, tens of thousands sentences

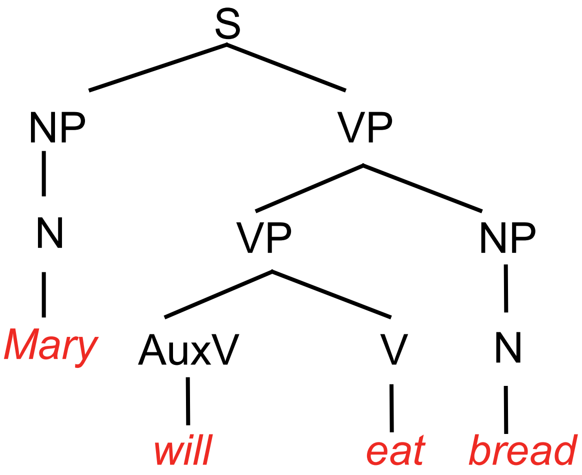
BUT:

- various data formats
- various user interface
- various annotation scenarios

Phrase structure trees:

- CFG-like trees
- non-terminals

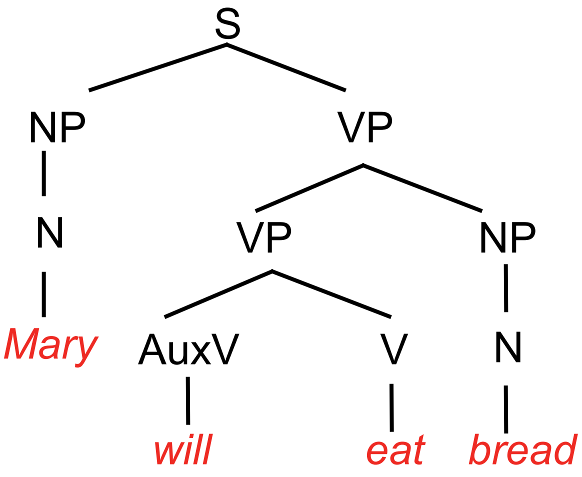
Mary will eat bread.



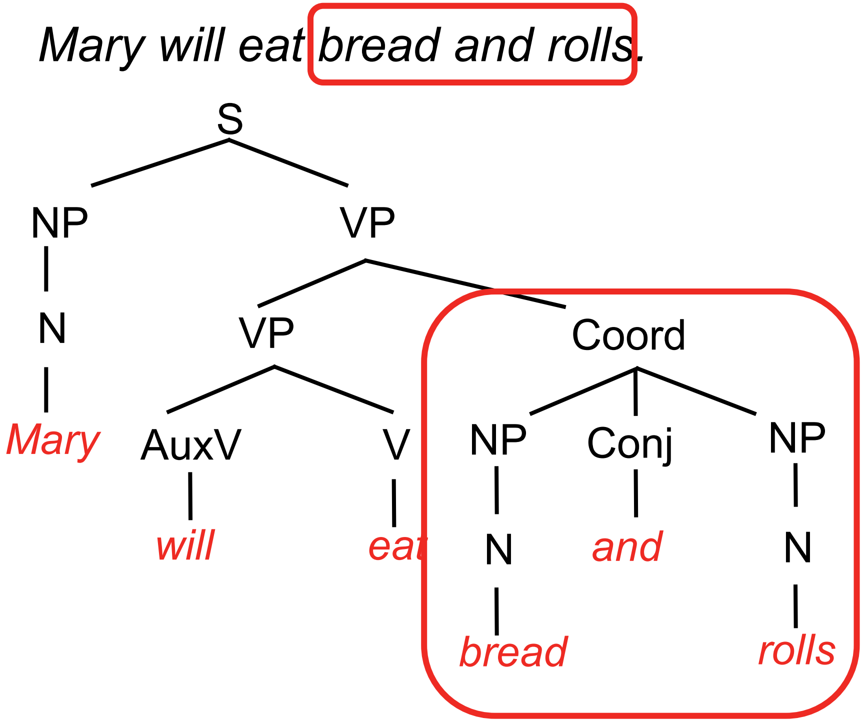
Phrase structure trees: Coordination

- CFG-like trees
- non-terminals

Mary will eat bread.



Mary will eat bread and rolls.

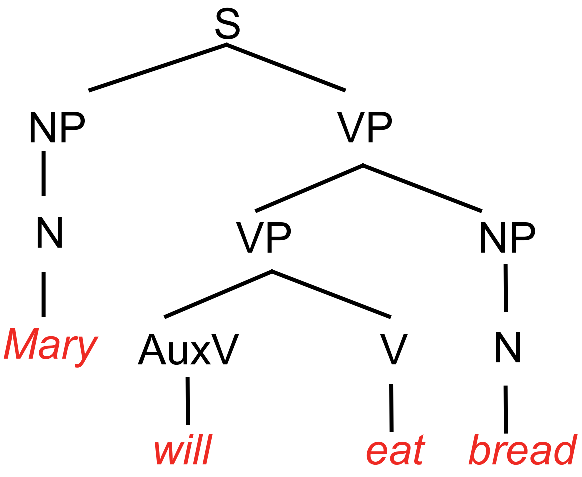


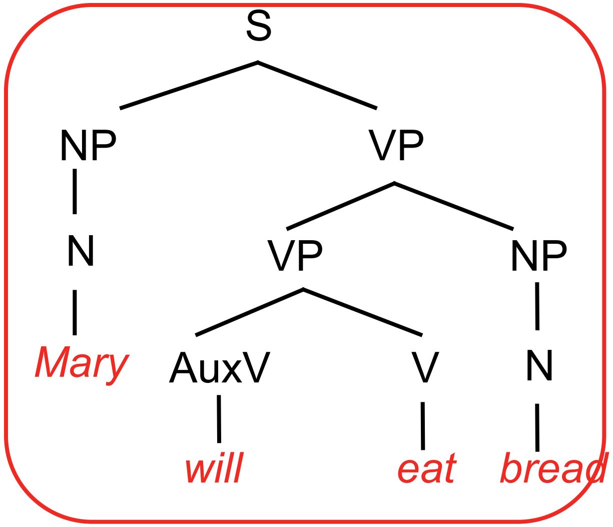
Phrase structure trees: Word order

- CFG-like trees
- non-terminals

What will Mary eat?

Mary will eat bread.

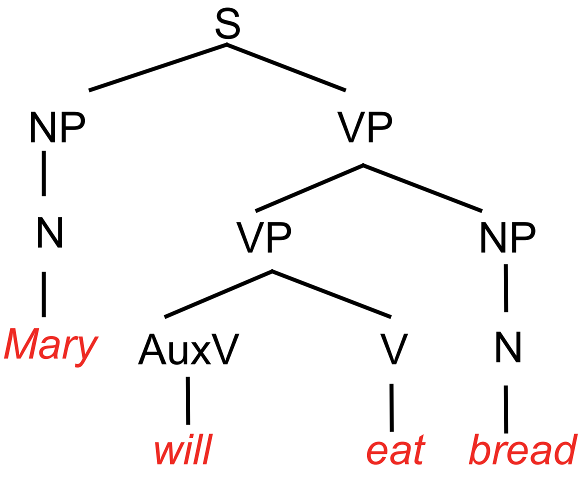




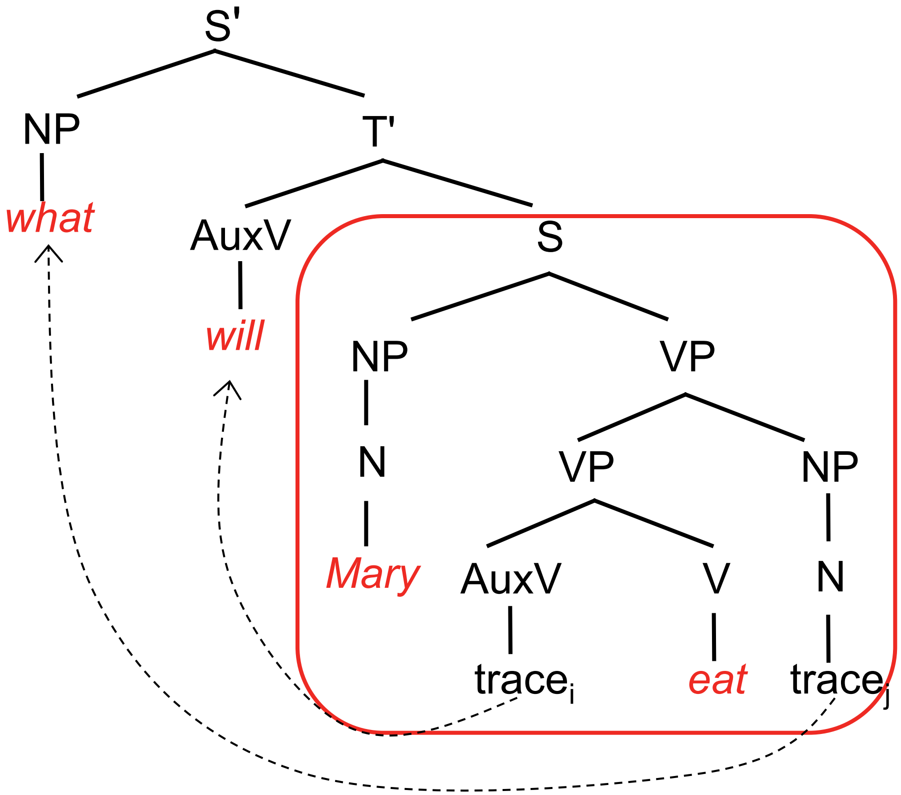
Phrase structure trees: Word order

- CFG-like trees
- non-terminals

Mary will eat bread.



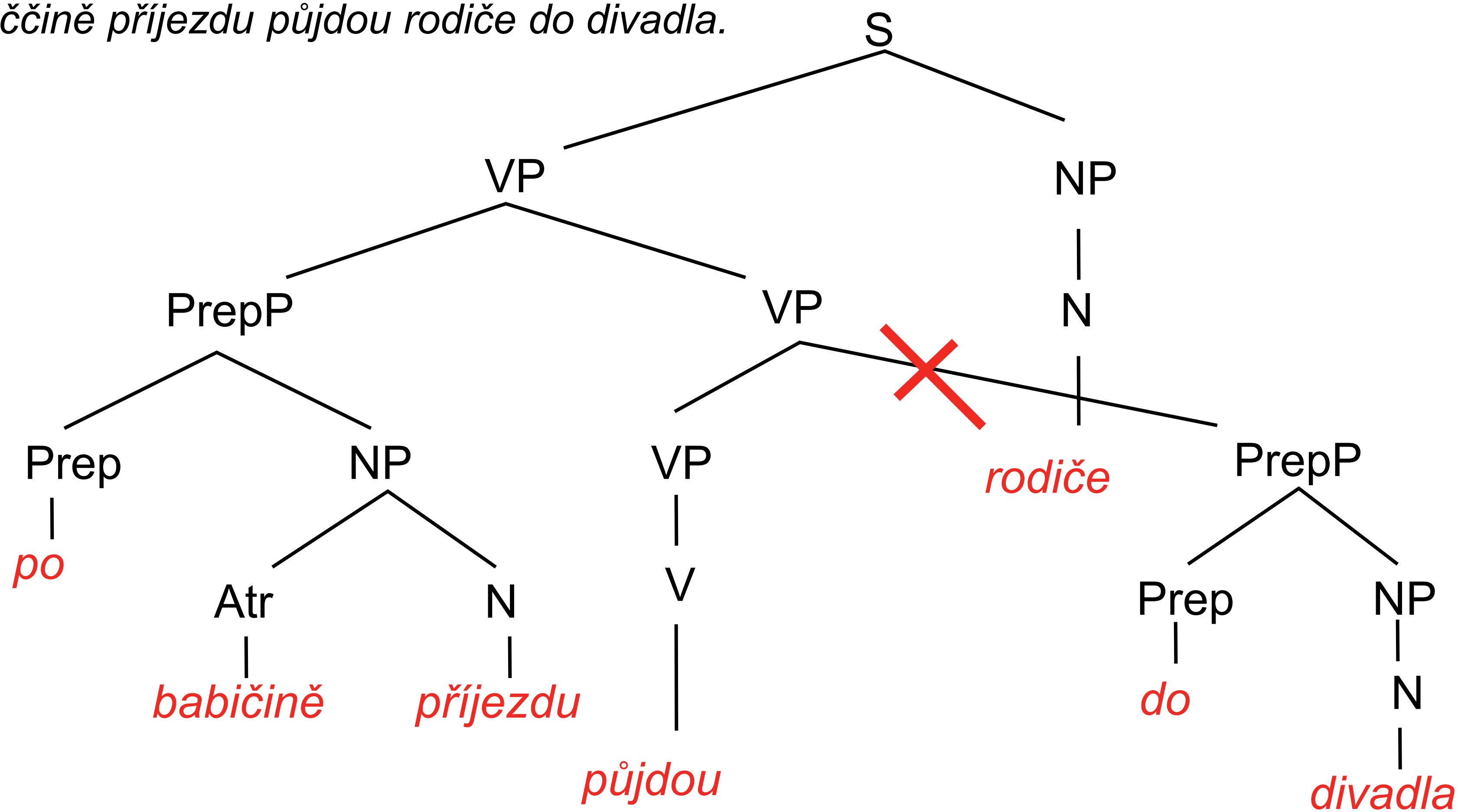
What will Mary eat?



Phrase structure trees: Word order

discontinuous ‘phrases’:

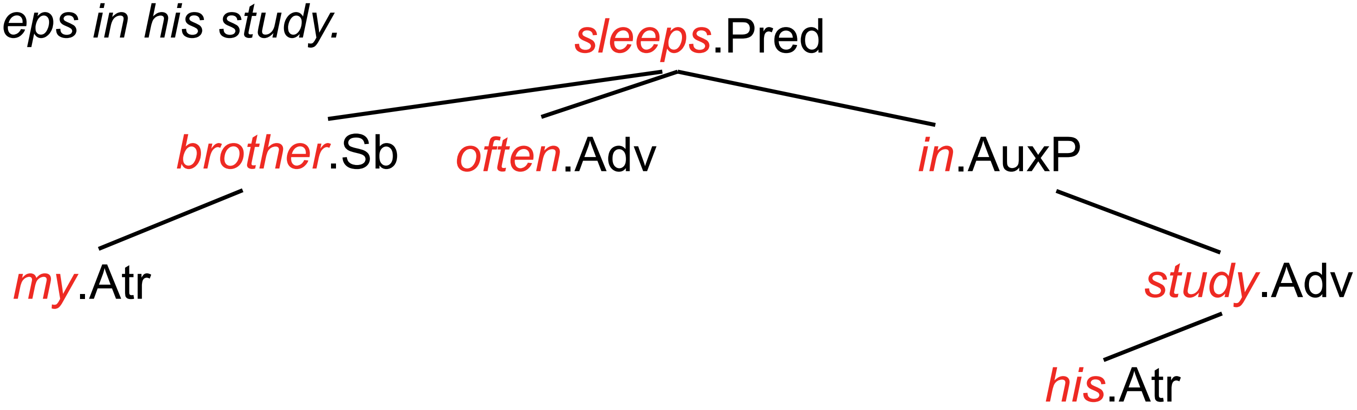
Po babiččině příjezdu půjdou rodiče do divadla.



Dependency trees: Word order

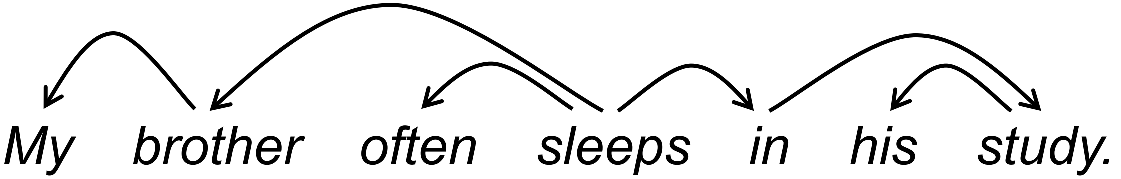
- lexicalised (= no non-terminals)
- nodes ~ lexical, morphological and syntactic information

My brother often sleeps in his study.



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

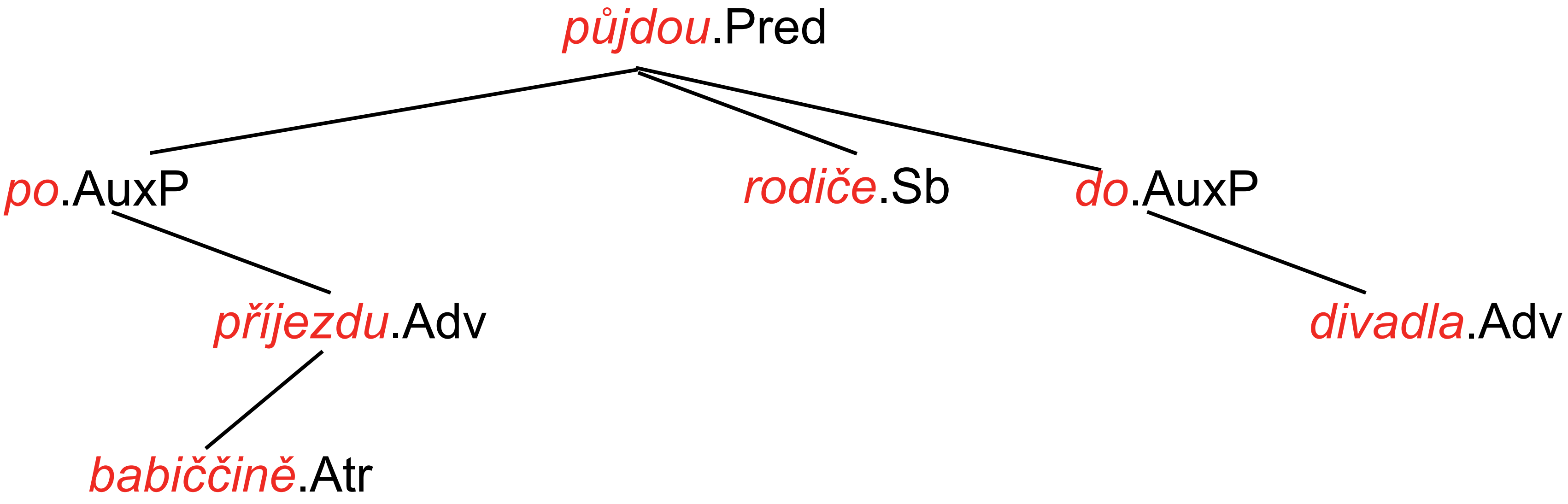
Igor Mel’čuk (1988) *Dependency Syntax: Theory and Practice.* State University of New York Press.



Dependency trees: Word order

discontinuous ‘phrases’:

Po babiččině příjezdu půjdou rodiče do divadla.



půjdou.Pred

po.AuxP

rodiče.Sb

do.AuxP

příjezdu.Adv

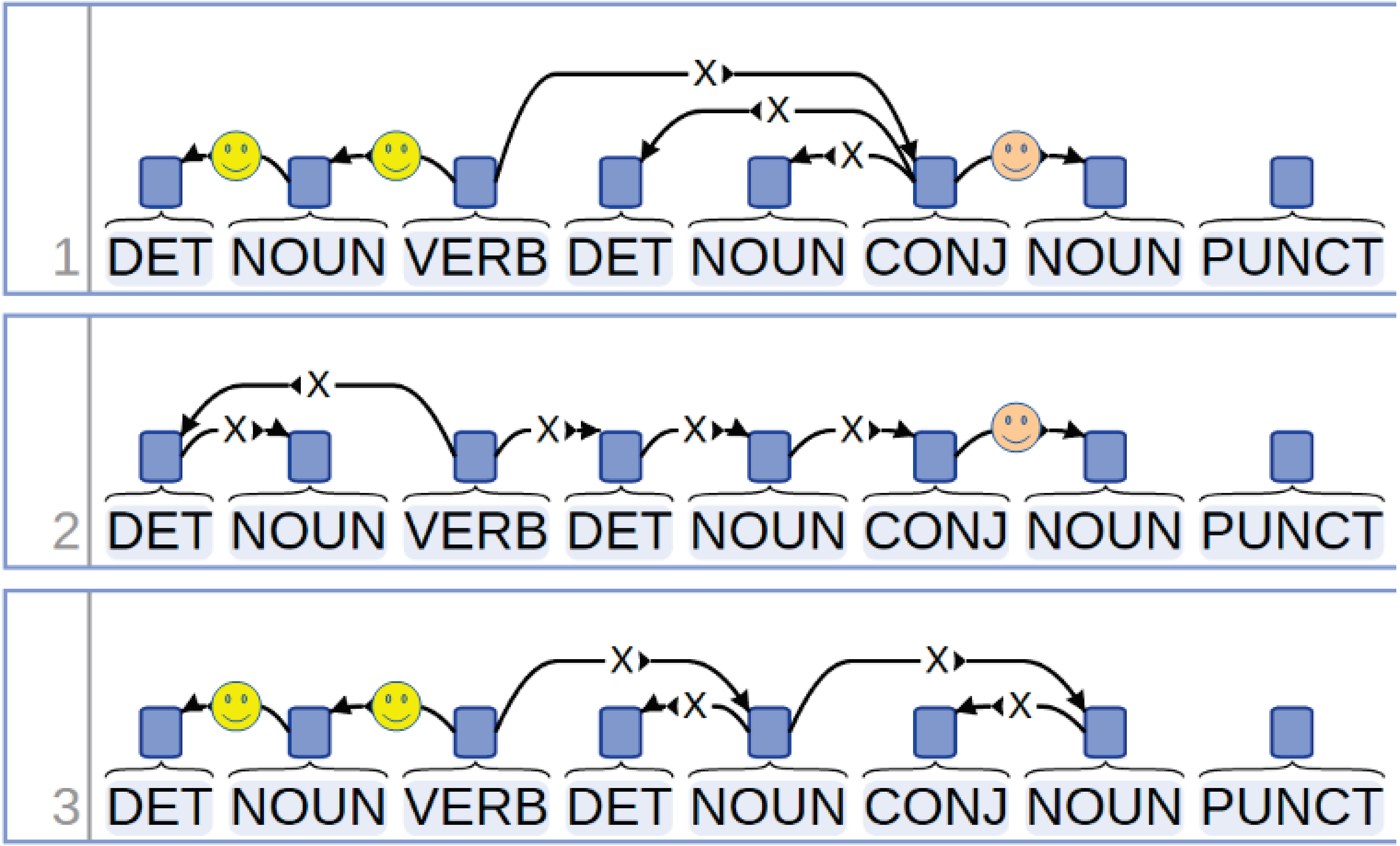
divadla.Adv

babiččině.Atr

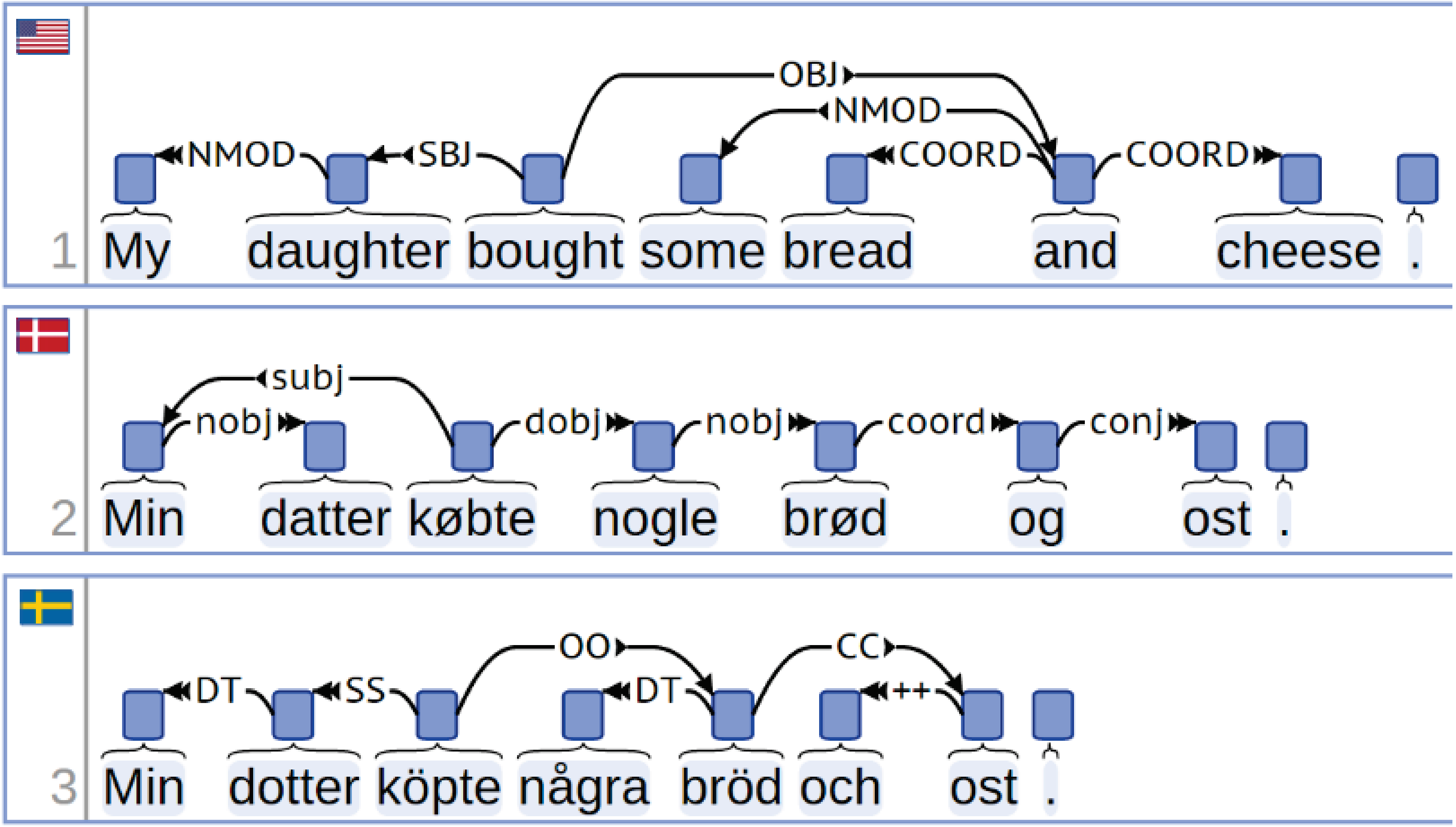
Dependency trees: Coordination

??? ‘standard’ ???

Why Treebanks Standardization?


1 DET NOUN VERB DET NOUN CONJ NOUN PUNCT
2 DET NOUN VERB DET NOUN CONJ NOUN PUNCT
3 DET NOUN VERB DET NOUN CONJ NOUN PUNCT
X

Why Treebanks Standardization?



1 My daughter bought some bread and cheese .
OBJ
NMOD
NMOD
SBJ
COORD
COORD
2 Min datter købte nogle brød og ost .
subj
nobj
dobj
nobj
coord
conj
3 Min dotter köpte några bröd och ost .
OO
CC
DT
SS
DT
++

Treebanks Standardization !!

- Prague Dependencies
    - from sixties/seventies (Nebeský. Sgall, Plátek)
    - as implemented in Prague Dependency Treebank (Hajič et al.) (nineties → now)
- Stanford Dependencies
- Google Universal Dependencies
- Stanford Universal Dependencies
- Google Universal Tags
- Universal Dependencies
  (Nivre, Zeman et al., from 2014)
- former alternatives

Prague Dependency Treebank (PDT)

PDT 2.0 (Hajič et al, 2006) and its upgrades
http://ufal.mff.cuni.cz/prague-dependency-treebank

- dependency relations
  - governing/modified unit (head) – dependent/modifying unit (modifier
  - criterion: possible reduction

  … dependent member of the pair may be deleted
  while the distributional properties are preserved (→ correctness is preserved)
- non-dependency relations
  - coordination as a core type
  - apposition
- other types
  (function words – auxiliaries, prepositions, punctuatiom, …)

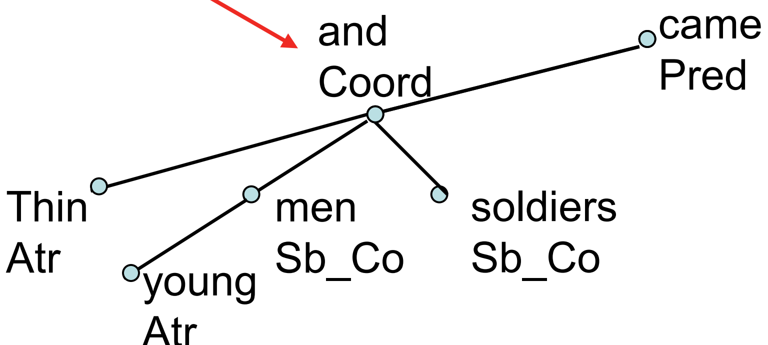
PDT: Coordination

as implemented in PML (Pajas, Štěpánek, 2005 → …)

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

specific types of nodes and edges:

- *connecting node* … Afun (a-layer) or nodetype + functor (t-layer)
  (= node for coordinating / appositing conjunction)



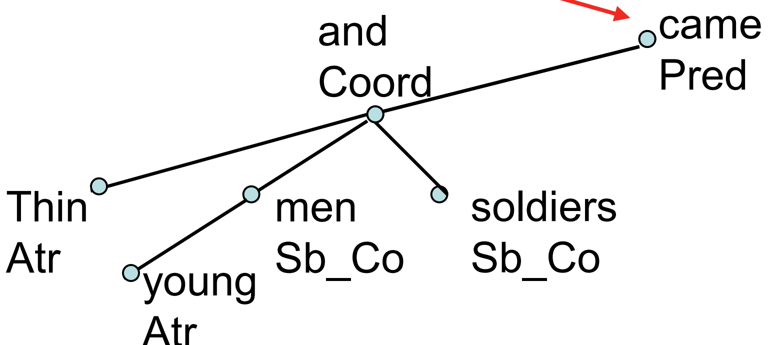
PDT: Coordination

as implemented in PML (Pajas, Štěpánek, 2005 → …)

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

specific types of nodes and edges:

- *connecting node* … Afun (a-layer) or nodetype + functor (t-layer)
  (= node for coordinating / appositing conjunction)
- *effective parent*
  (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')



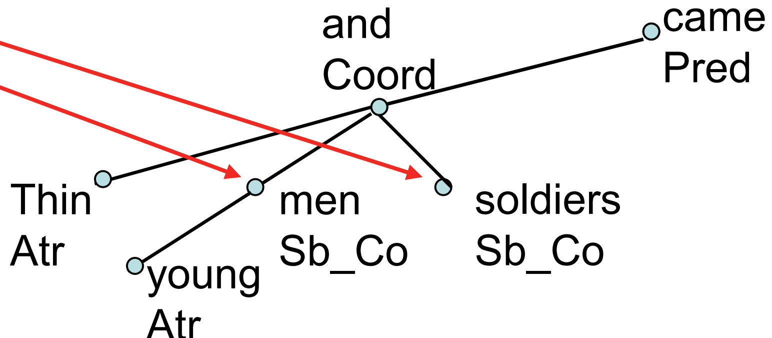
PDT: Coordination

as implemented in PML (Pajas, Štěpánek, 2005 → …)

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

specific types of nodes and edges:

- *connecting node* … Afun (a-layer) or nodetype + functor (t-layer)
  (= node for coordinating / appositing conjunction or punctuation)
- *effective parent*
  (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
- *members of a connecting construction* … is_member
  (= nodes that are coordinated / are in apposition)



PDT: Coordination

as implemented in PML (Pajas, Štěpánek, 2005 → …)

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

specific types of nodes and edges:

- *connecting node* … Afun (a-layer) or nodetype + functor (t-layer)

  (= node for coordinating / appositing conjunction or punctuation)
- *effective parent*

  (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
- *members of a connecting construction* … is_member

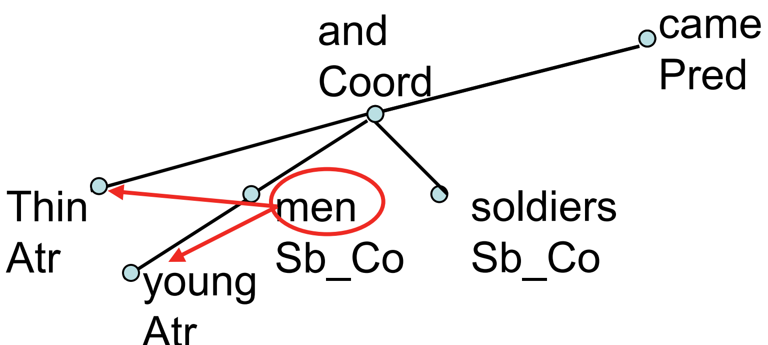
  (= nodes that are coordinated / are in apposition)
- *effective child(ren)*

  ('linguistic dependency')

  e.g., *men – young; men – thin*

  soldiers – thin

  came – men; came – soldiers



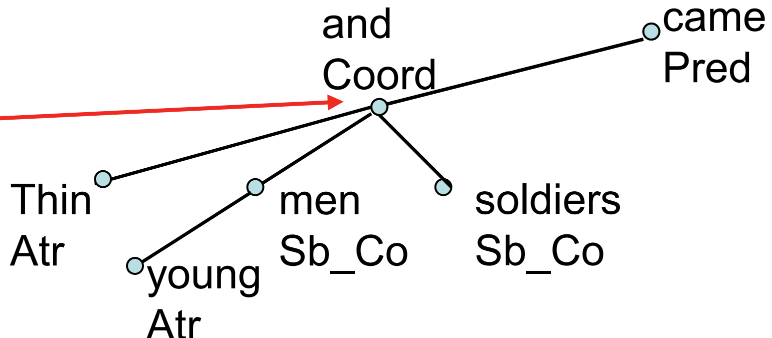
PDT: Coordination

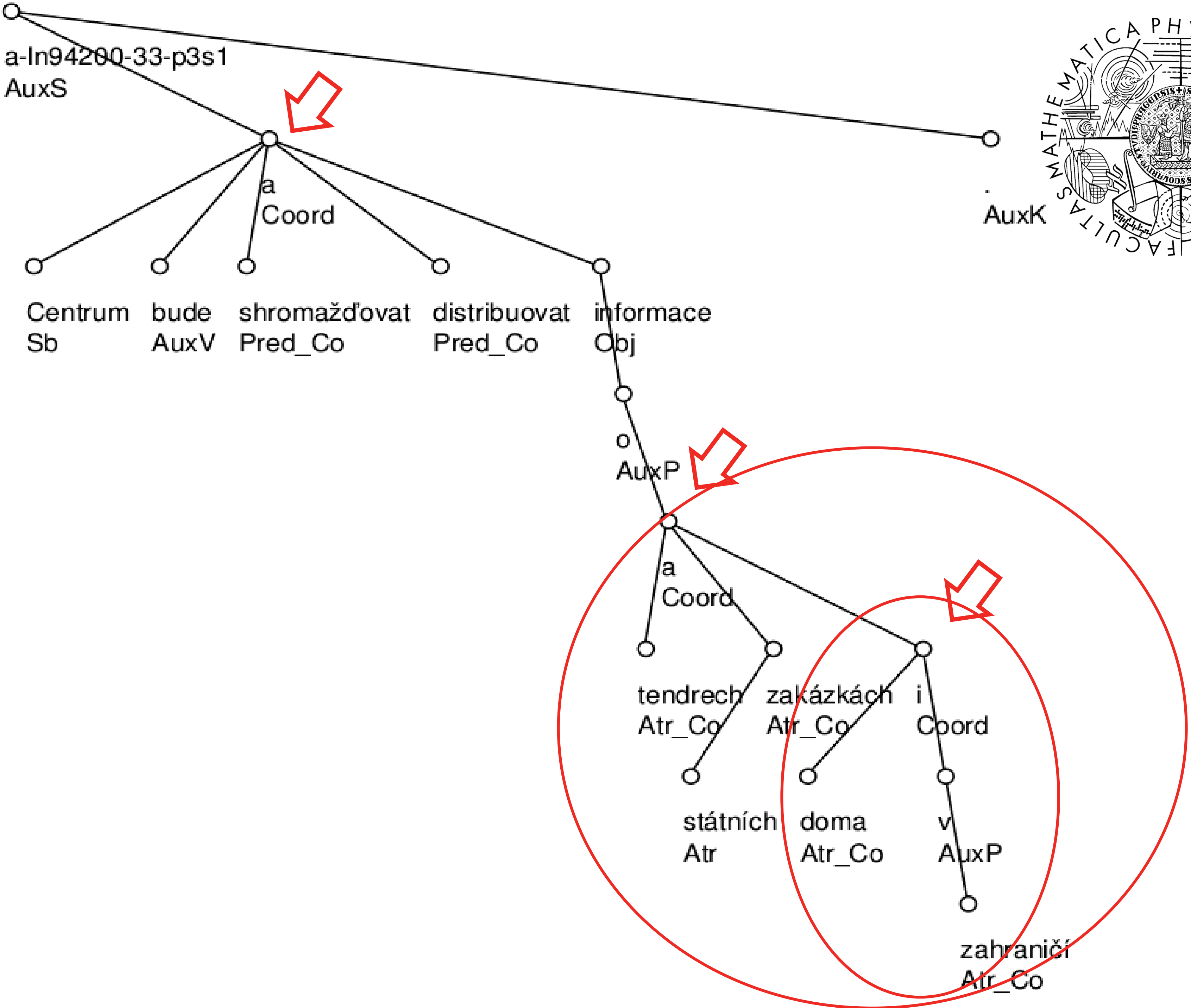
'connecting' constructions ~ coordination, apposition (, OPER)

specific types of nodes and edges:

- *connecting node* … Afun (a-layer) or nodetype + functor (t-layer)
  (= node for coordinating / appositing conjunction or punctuation)
- *effective parent*
  (= node for governing node, i.e. node modified by the whole construction, 'linguistic parent')
- *members of a connecting construction* … is_member
  (= nodes that are coordinated / are in apposition)
- *effective child(ren)*
  (‘linguistic dependency’)

- *‘pass-through’ nodes*
  ~ conjunctions, prepositions

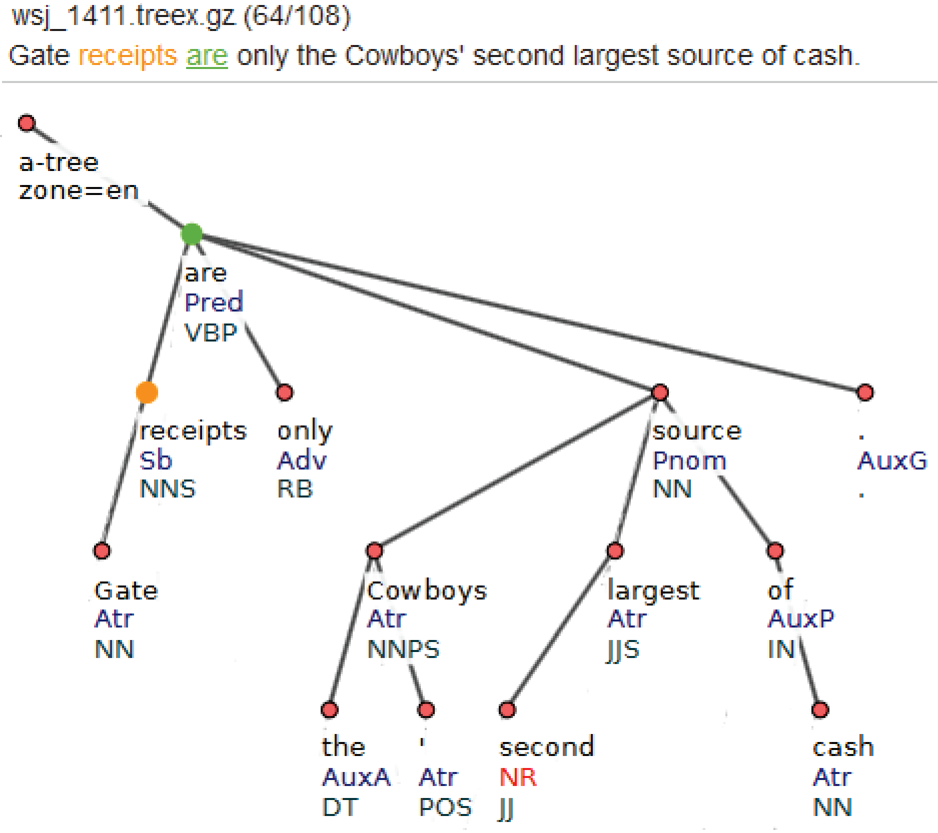



a-ln94200-33-p3s1
AuxS
a
Coord
.
AuxK
Centrum
Sb
bude
AuxV
shromažďovat
Pred_Co
distribuovat
Pred_Co
informace
Obj
o
AuxP
a
Coord
tendrech
Atr_Co
zakázkách
Atr_Co
i
Coord
státních
Atr
doma
Atr_Co
v
AuxP
zahraničí
Atr_Co

HamleDT: HArmonized Multi-LanguagE Dependency Treebank

- 42 treebanks for 36 languages (Zeman et al., from 2012)
- common format
  - based on Prague Dependency Treebank scenario
    - minor changes
  - (semi)automatic conversion from original treebanks
  - freely available whenever possible (license constraints)
  - http://ufal.mff.cuni.cz/hamledt

HamleDT:



wsj_1411.treex.gz (64/108)

Gate receipts are only the Cowboys' second largest source of cash.

Universal Dependencies (UD)

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards
- Caveats:
    - not a new linguistic theory – but linguistically informed and relevant
    - not an ideal parsing representation – but useful for comparative evaluation
    - not the ultimate annotation scheme – but a lightweight lingua franca

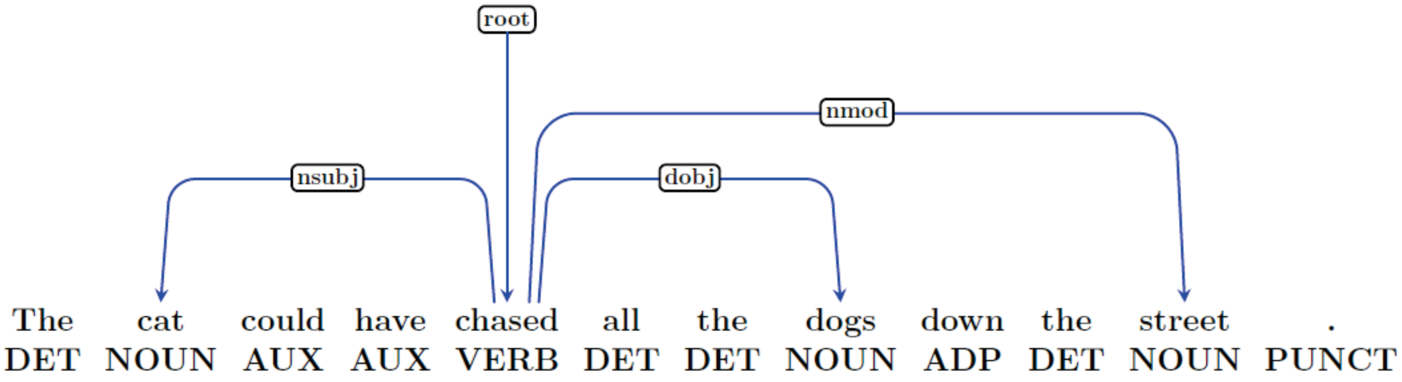
(Slides stolen from Daniel Zeman, Joakim Nivre)

UD Syntax

Basic principles:

- content words are related by dependency relations

Why: to stress language similarities


root
nmod
nsubj
dobj
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT

(Slides based on slides by Daniel Zeman, Joakim Nivre)

UD Syntax

Basic principles:

- content words are related by dependency relations

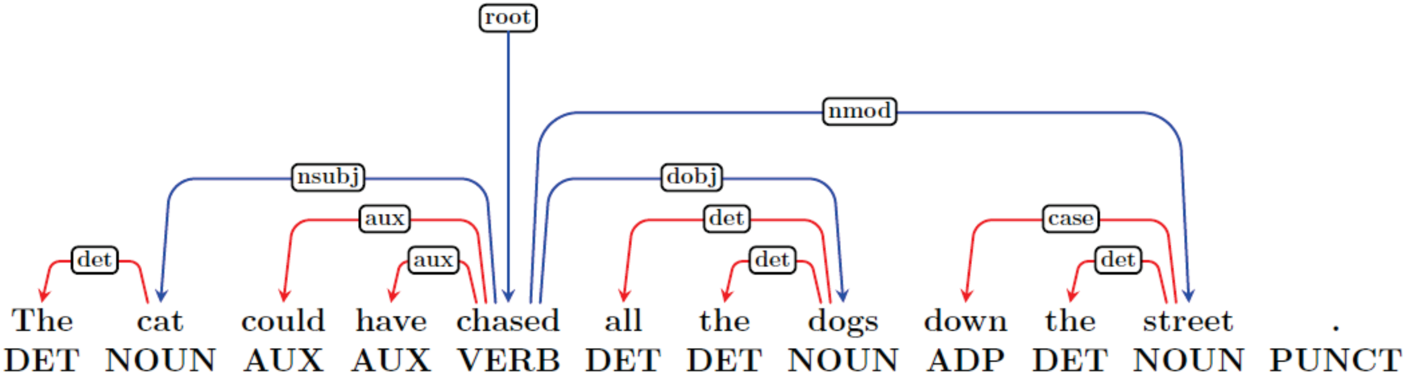
  Why: to stress language similarities

- function words attached to closest content word

  Why: as languages differ wrt. function words,

  e.g. preposition/less phrases

  Petr dal dárek Marii. – Peter gave the gift to Mary.



(Slides based on slides by Daniel Zeman, Joakim Nivre)

UD Syntax

Basic principles:

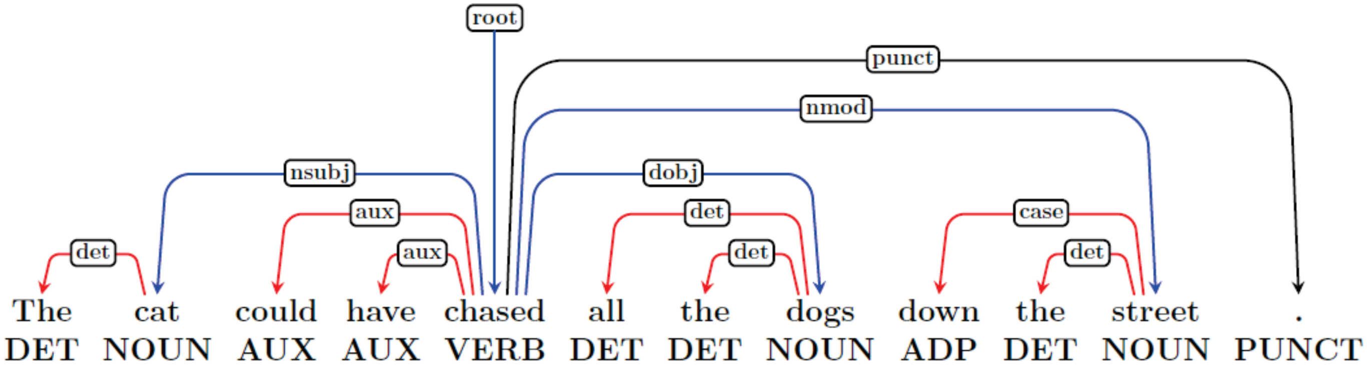
- content words are related by dependency relations

  Why: to stress language similarities

- function words attached to closest content word

  Why: as languages differ wrt. function words,

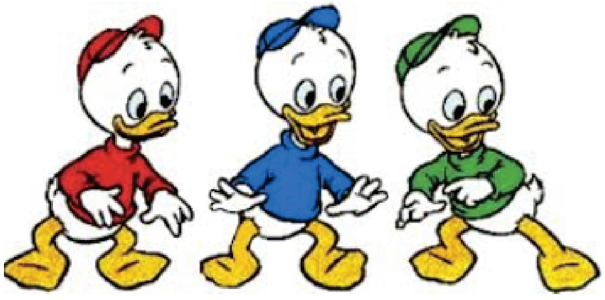
- punctuation attached to head of phrase or clause

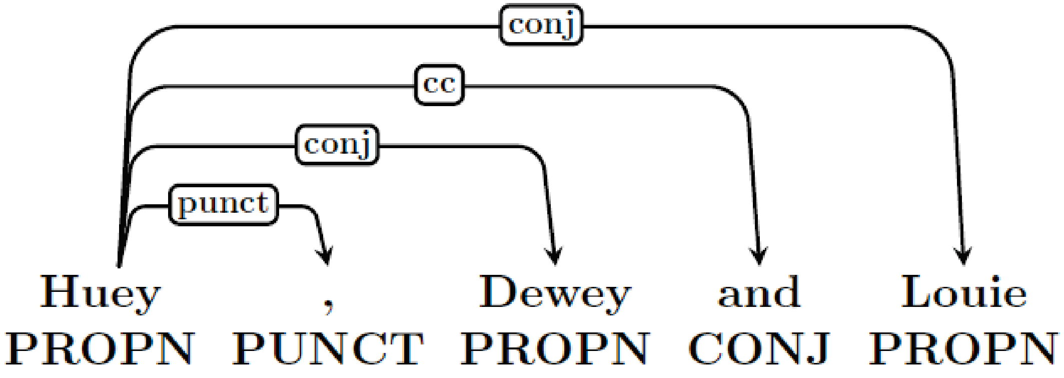

root
punct
nmod
nsubj
dobj
aux
det
det
case
det
aux
det
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT

(Slides based on slides by Daniel Zeman, Joakim Nivre)

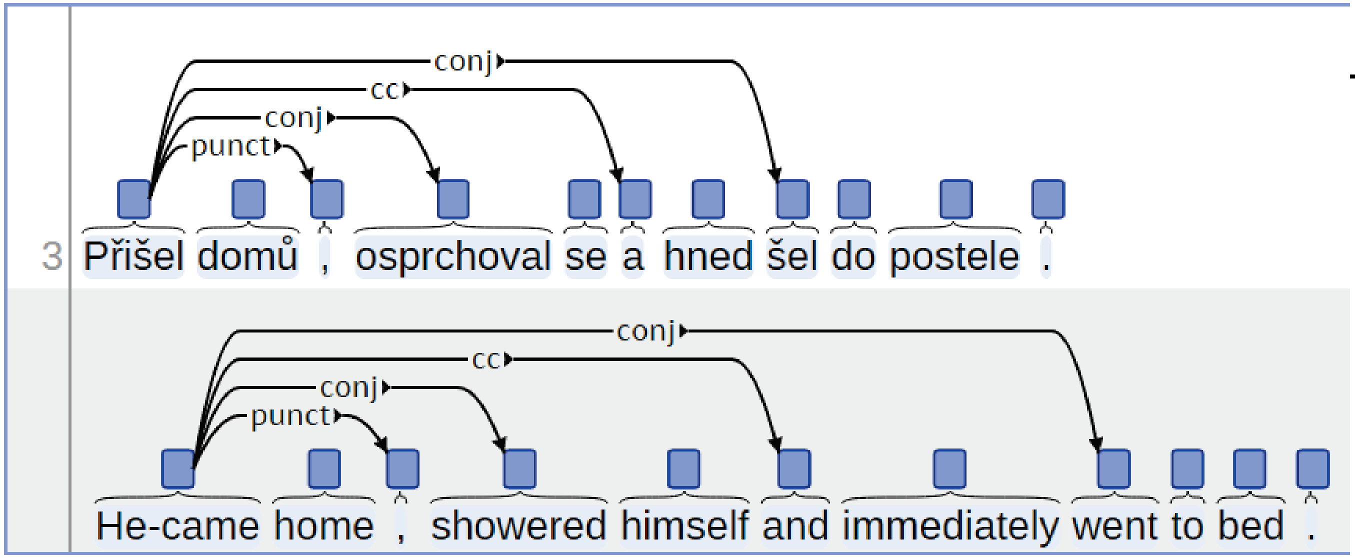
UD Syntax: Coordination

- Coordinate structures are headed by the first conjunct
  - subsequent conjuncts depend on it via the conj relation
  - conjunctions depend on it via the cc relation
  - punctuation marks depend on it via the punct relation

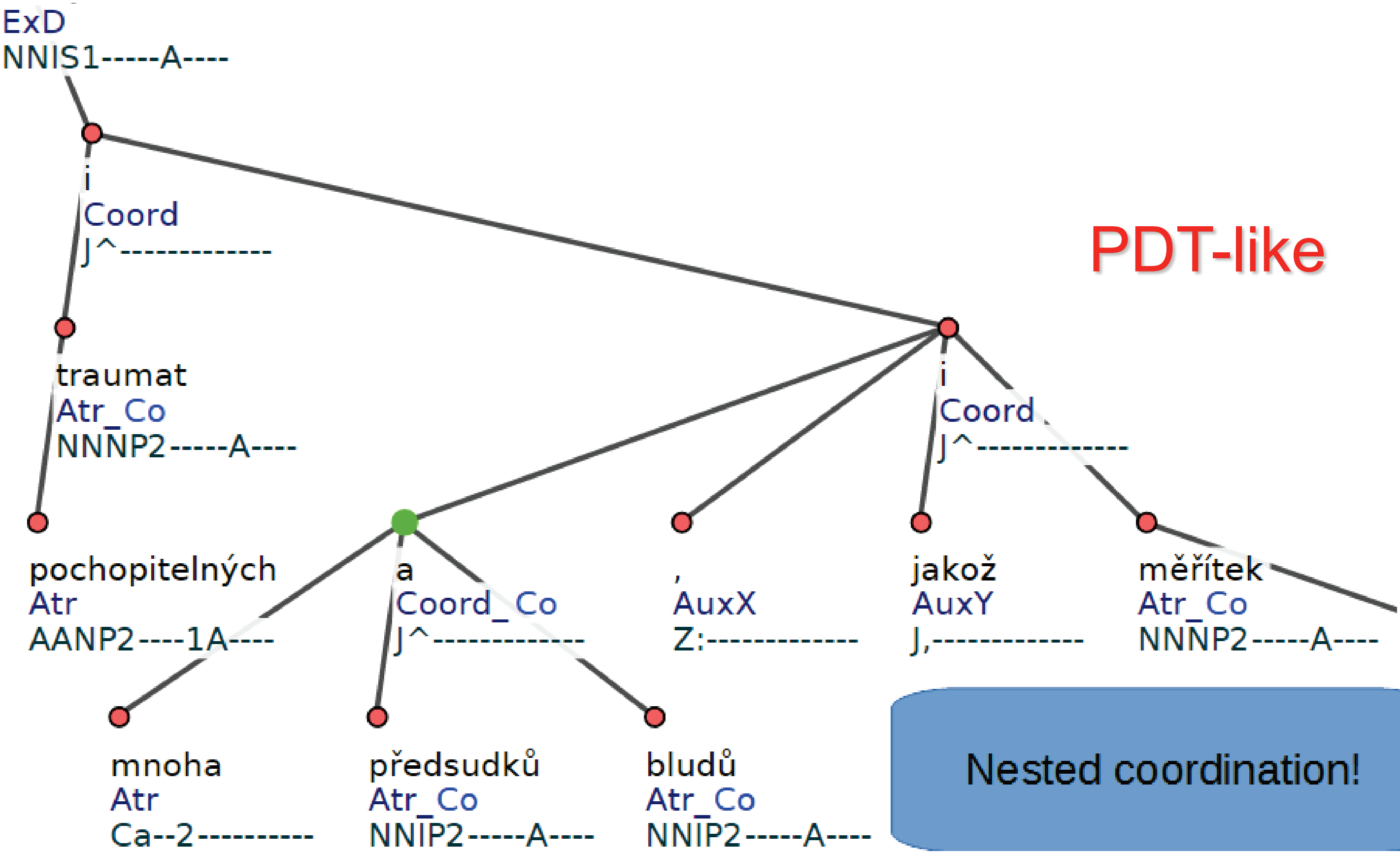




(Slides stolen from Daniel Zeman)


conj
cc
conj
punct
3 Přišel domů , osprchoval se a hned šel do postele .
conj
cc
conj
punct
He-came home , showered himself and immediately went to bed .

UD-like

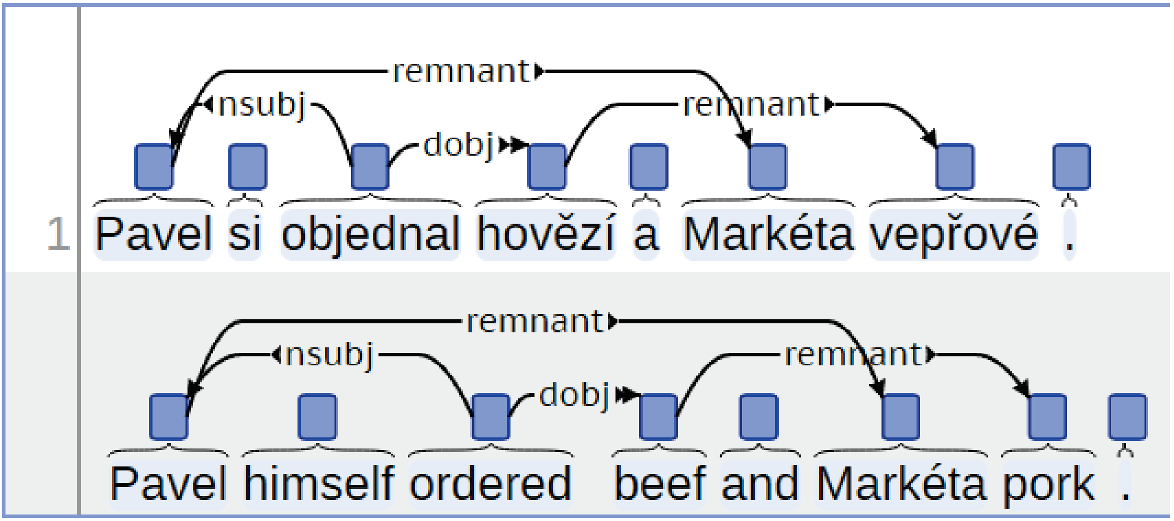


PDT-like

Shared dependent!

Nested coordination!

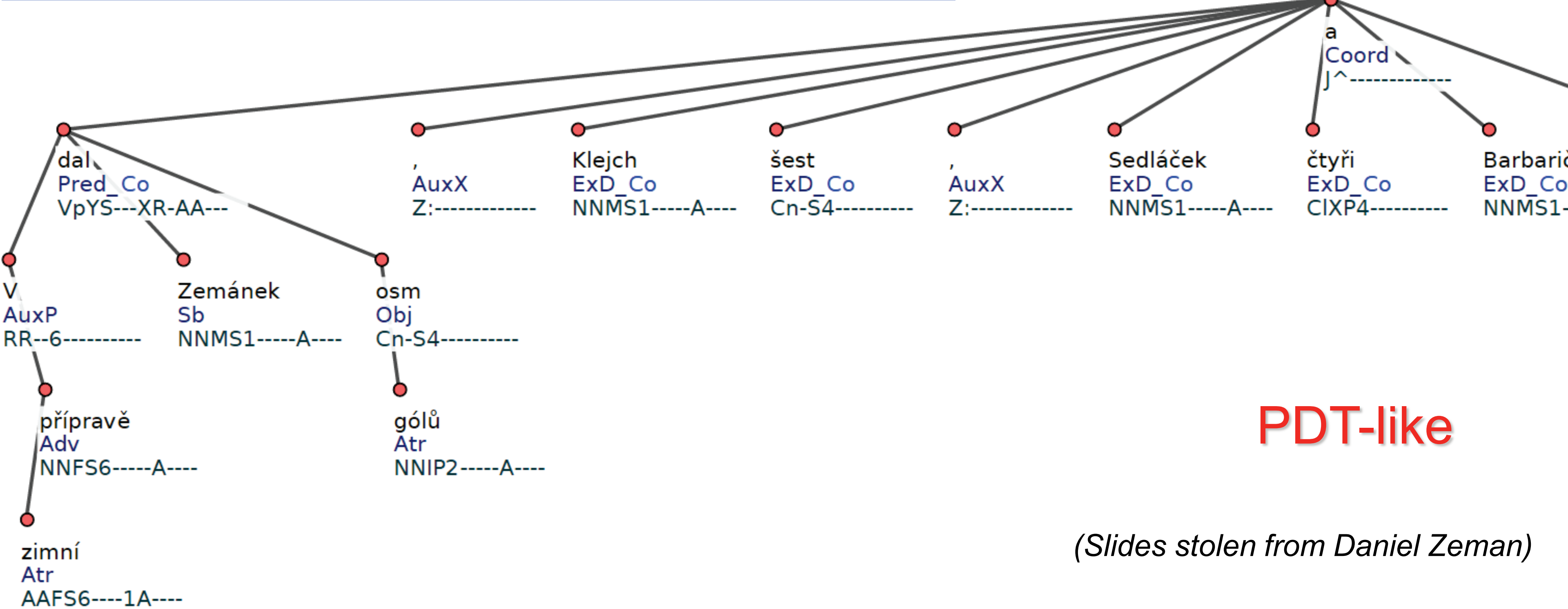
Coordination with ellipses



1 Pavel si objednal hovězí a Markéta vepřové .

Pavel himself ordered beef and Markéta pork .

UD-like



PDT-like

(Slides stolen from Daniel Zeman)

Universal Dependencies

- 2014-04: EACL Göteborg, kick-off meeting
- 2014-10: UD guidelines version 1
- 2015-01: released treebanks of 10 languages (UD 1.0)
- 2015-05: released treebanks of 18 languages (UD 1.1)
- 2015-11: next release

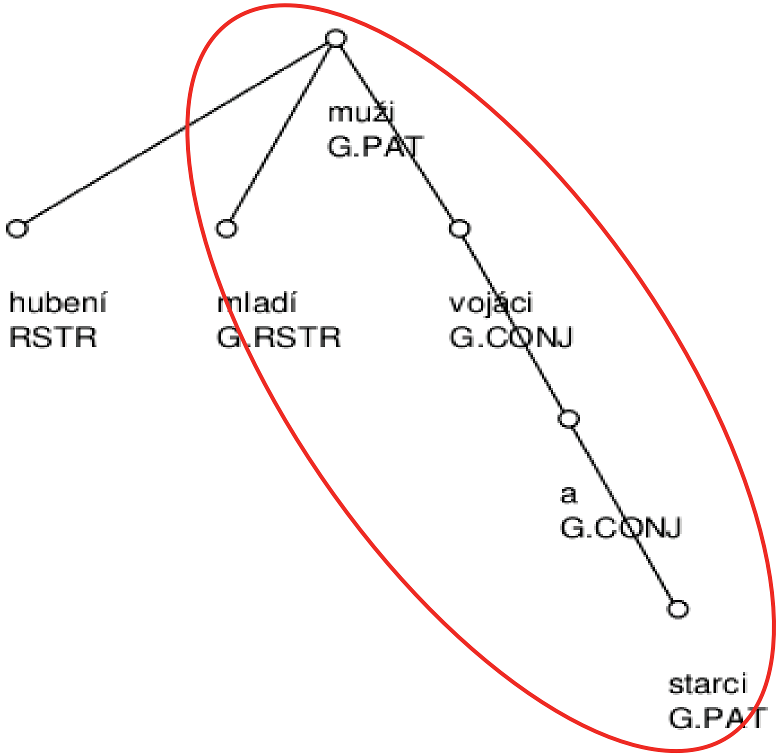
(Slides stolen from Daniel Zeman)

Alternative solution I

Mel'čuk (1988)

problems:

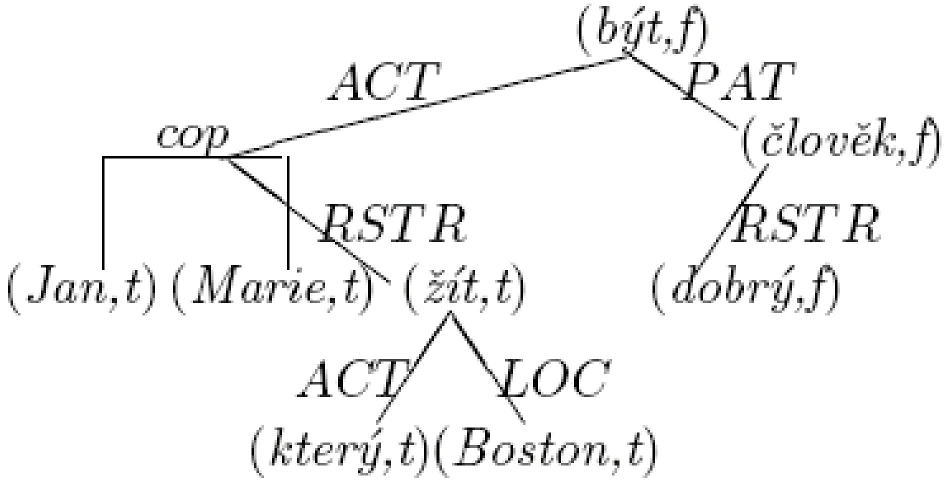
- shared modification vs. modification of a single member
- embedded coordinations



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

Alternative solution II

Petkevič (1995) … formal representation of FGD



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

Alternative solution III

… and many others