Rudolf Rosa
rosa@ufal.mff.cuni.cz

Syntaktická analýza vět přirozeného jazyka pomocí částečně řízených metod

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Státní doktorská zkouška, 15. června 2015

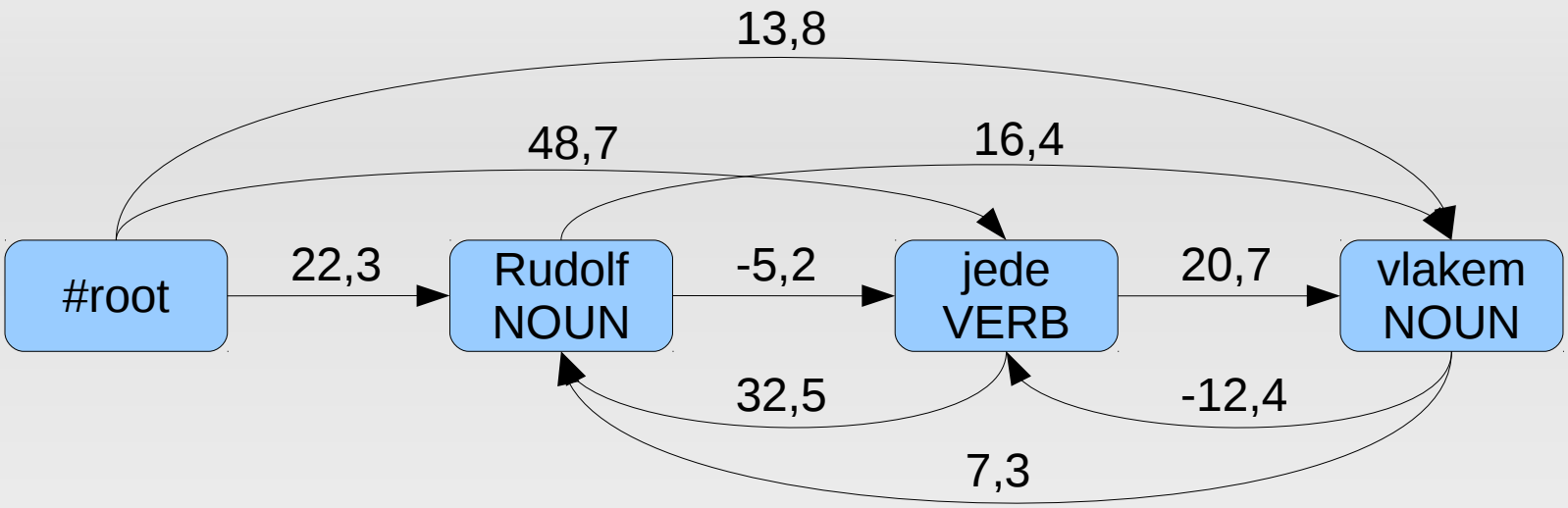
Osnova

- Úvod a motivace
- MSTParser a jeho delexikalizace
- Přenos delexikalizovaného parseru
  - s jedním zdrojem
  - s více zdroji
    - kombinace stromů
    - interpolace modelů
- Volba anotačního stylu pro parsing
- Výhled do budoucna: mezijazyčná lexikalizace

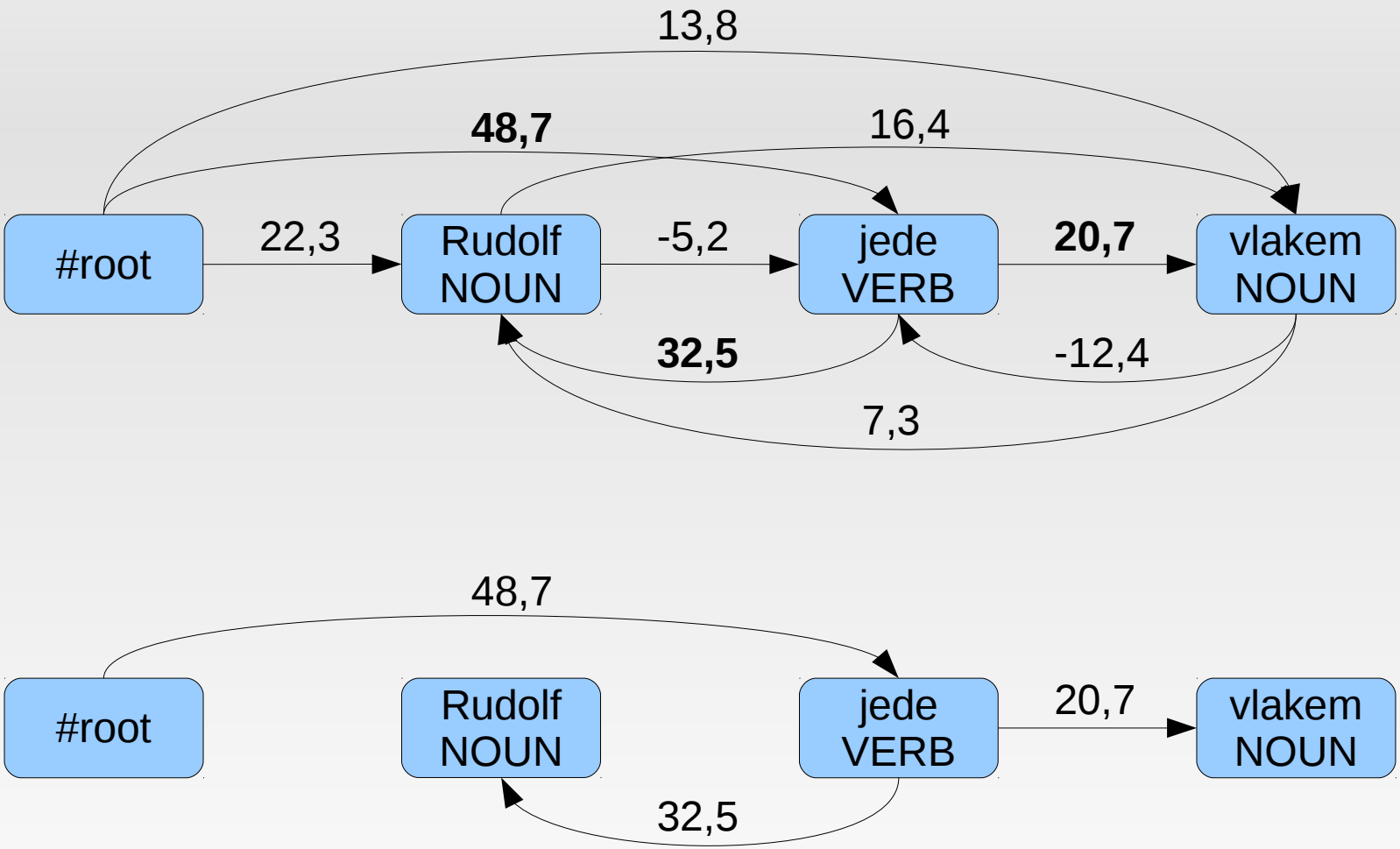
Částečně řízený parsing

- plně řízená syntaktická analýza vět (parsing)
  - vyžaduje trénovací data (treebank) nebo gramatiku
  - existuje řádově 100 treebanků (ruční práce)
  - existuje zhruba 7000 jazyků
  - nemluvě o různých doménách, vývoji jazyka...
- částečně řízený parsing
  - využít existující zdroje, nic nově neanotovat
    - treebanky pro jiné jazyky (HamleDT: 30 jazyků)
    - neanotovaná data (v mojí práci: otagovaná)

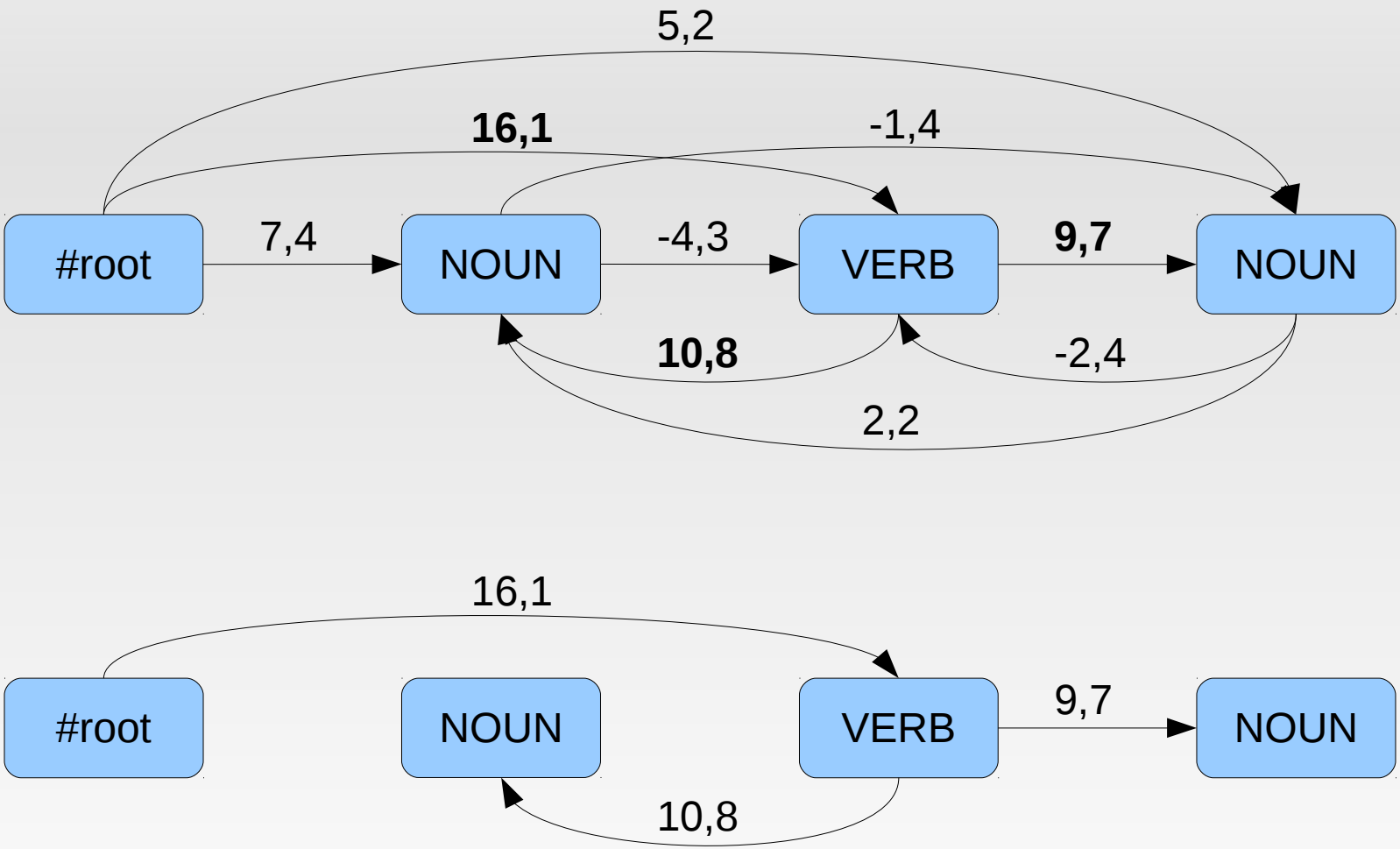
(Lexikalizovaný) MSTParser



(Lexikalizovaný) MSTParser



Delexikalizovaný MSTParser



Přenos delex. parseru s 1 zdrojem

- natrénovat delex parser na jazyku s treebankem
- aplikovat na jazyk bez treebanku (s taggerem)

Přenos delex. parseru s 1 zdrojem

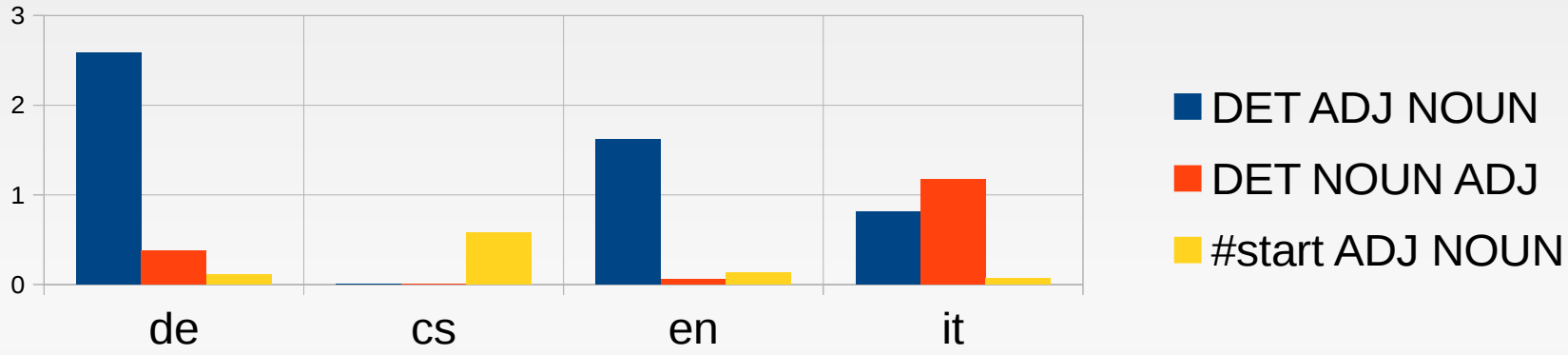
- natrénovat delex parser na jazyku s treebankem
- aplikovat na jazyk bez treebanku (s taggerem)
- jak vybrat zdrojový jazyk?

Přenos delex. parseru s 1 zdrojem

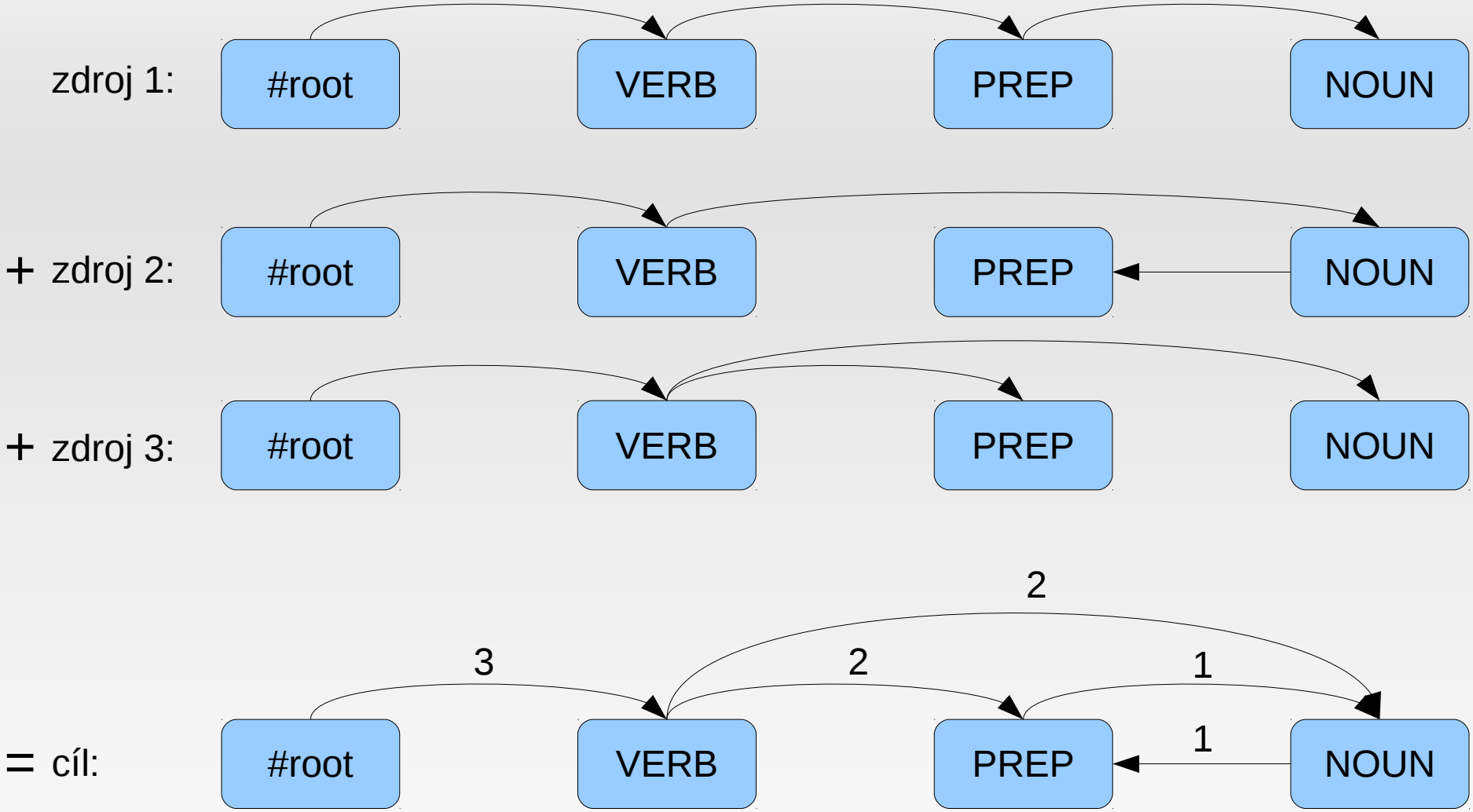
- natrénovat delex parser na jazyku s treebankem
- aplikovat na jazyk bez treebanku (s taggerem)
- jak vybrat zdrojový jazyk?
  - vybrat jazyk co nejpodobnější cílovému

Přenos delex. parseru s 1 zdrojem

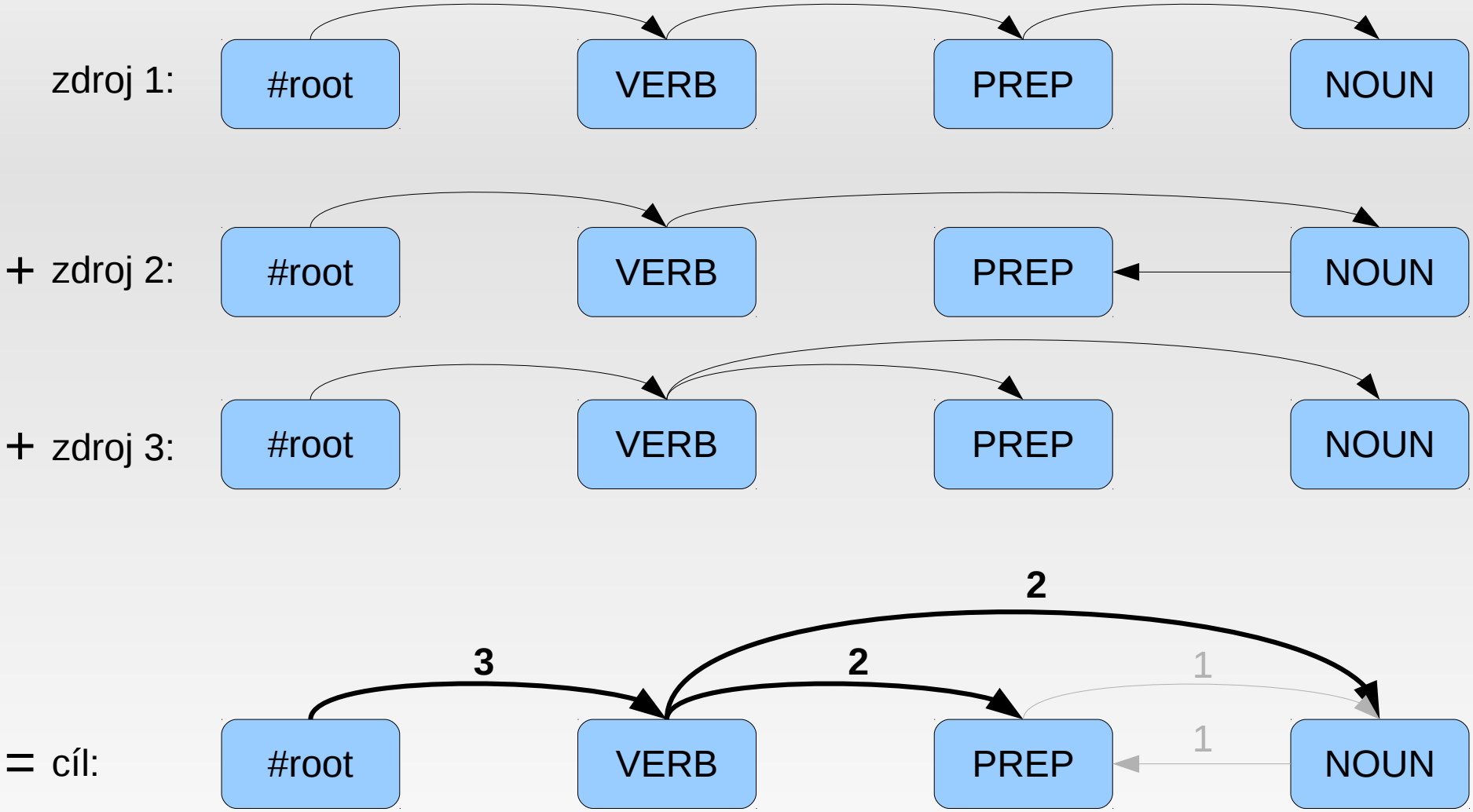
- natrénovat delex parser na jazyku s treebankem
- aplikovat na jazyk bez treebanku (s taggerem)
- jak vybrat zdrojový jazyk?
  - vybrat jazyk co nejpodobnější cílovému
  - KL_{cpos3}: distribuce trojic slovních druhů v korpusech



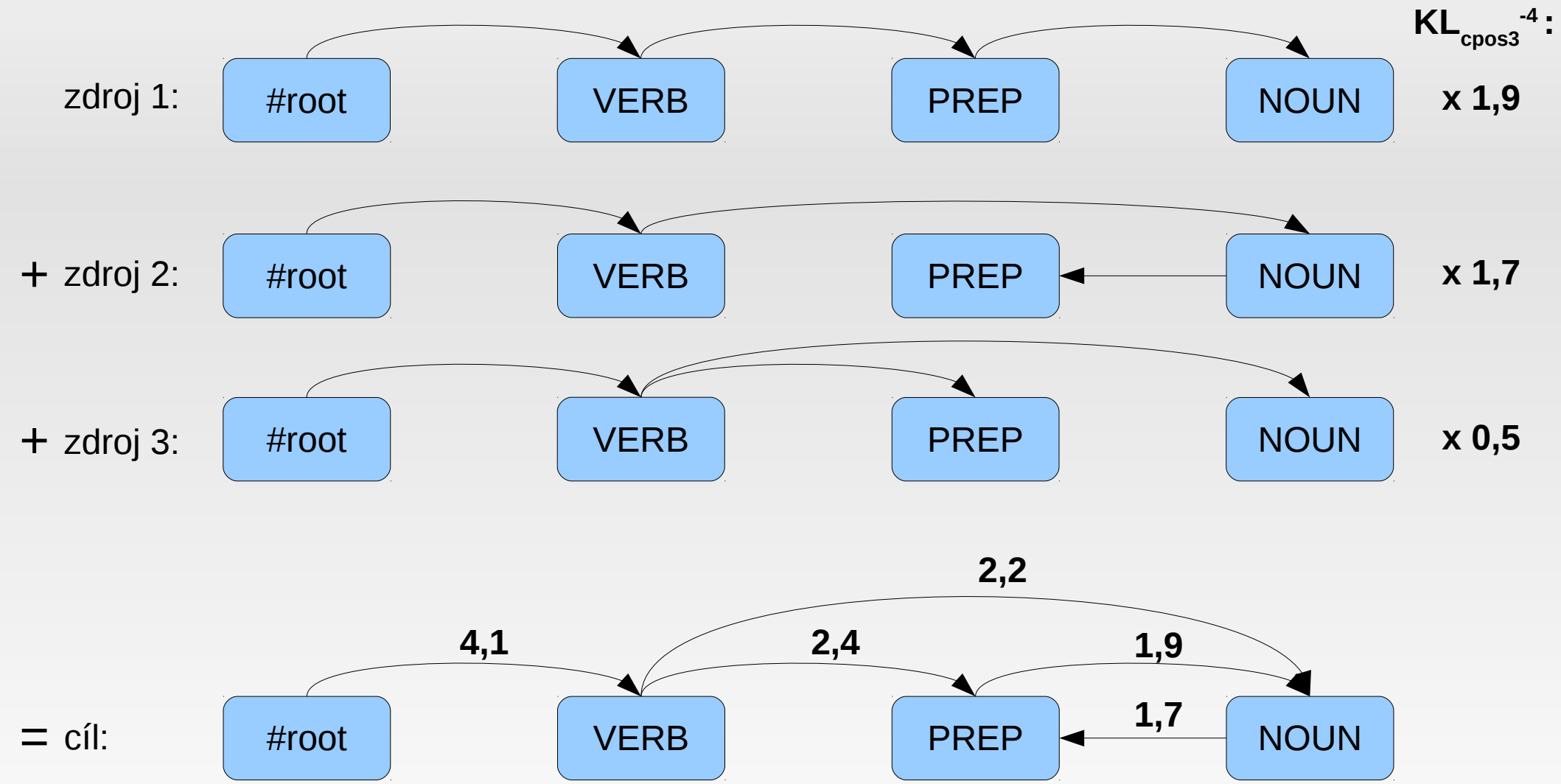
Více zdrojů: kombinace stromů



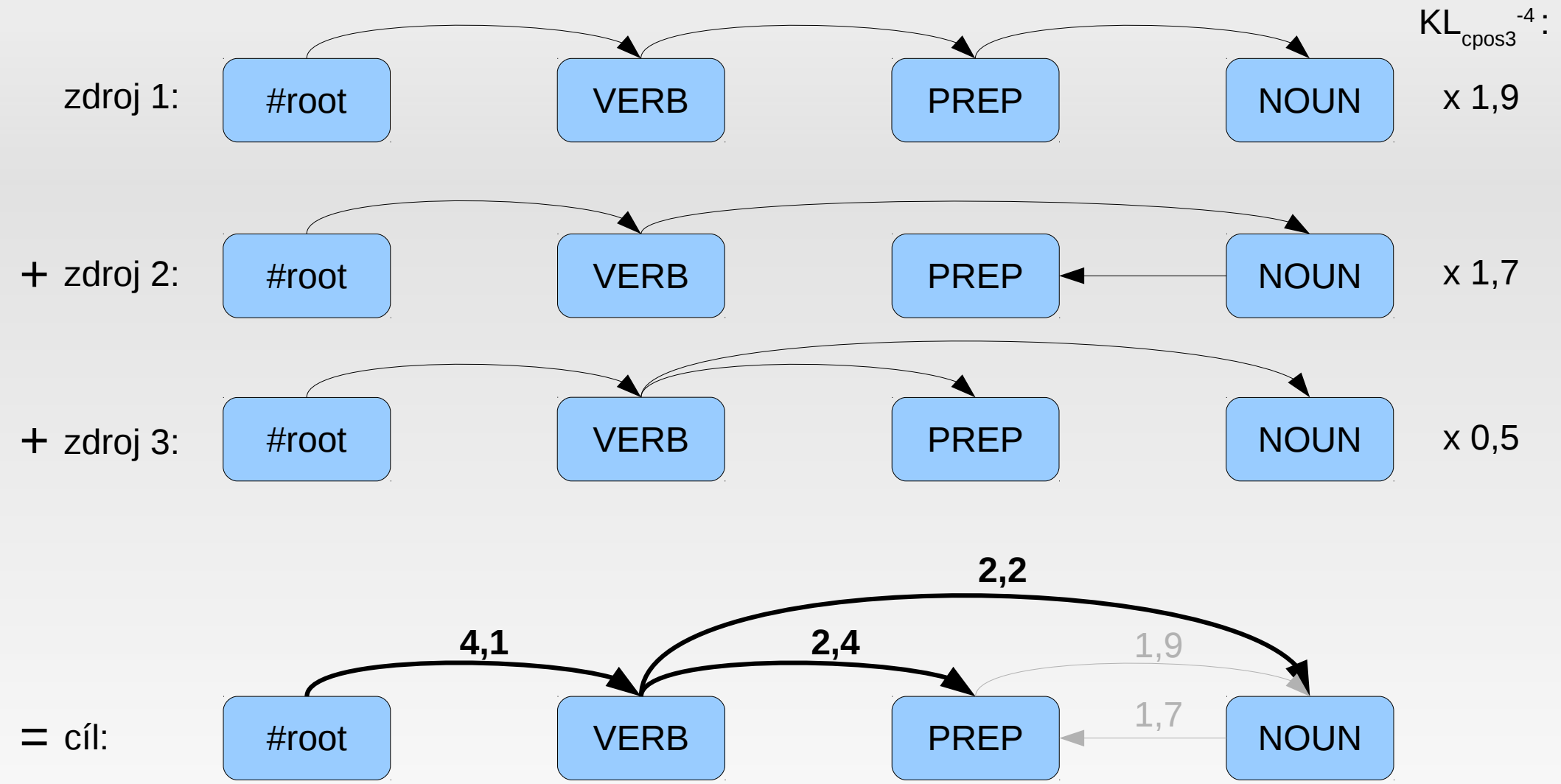
Více zdrojů: kombinace stromů



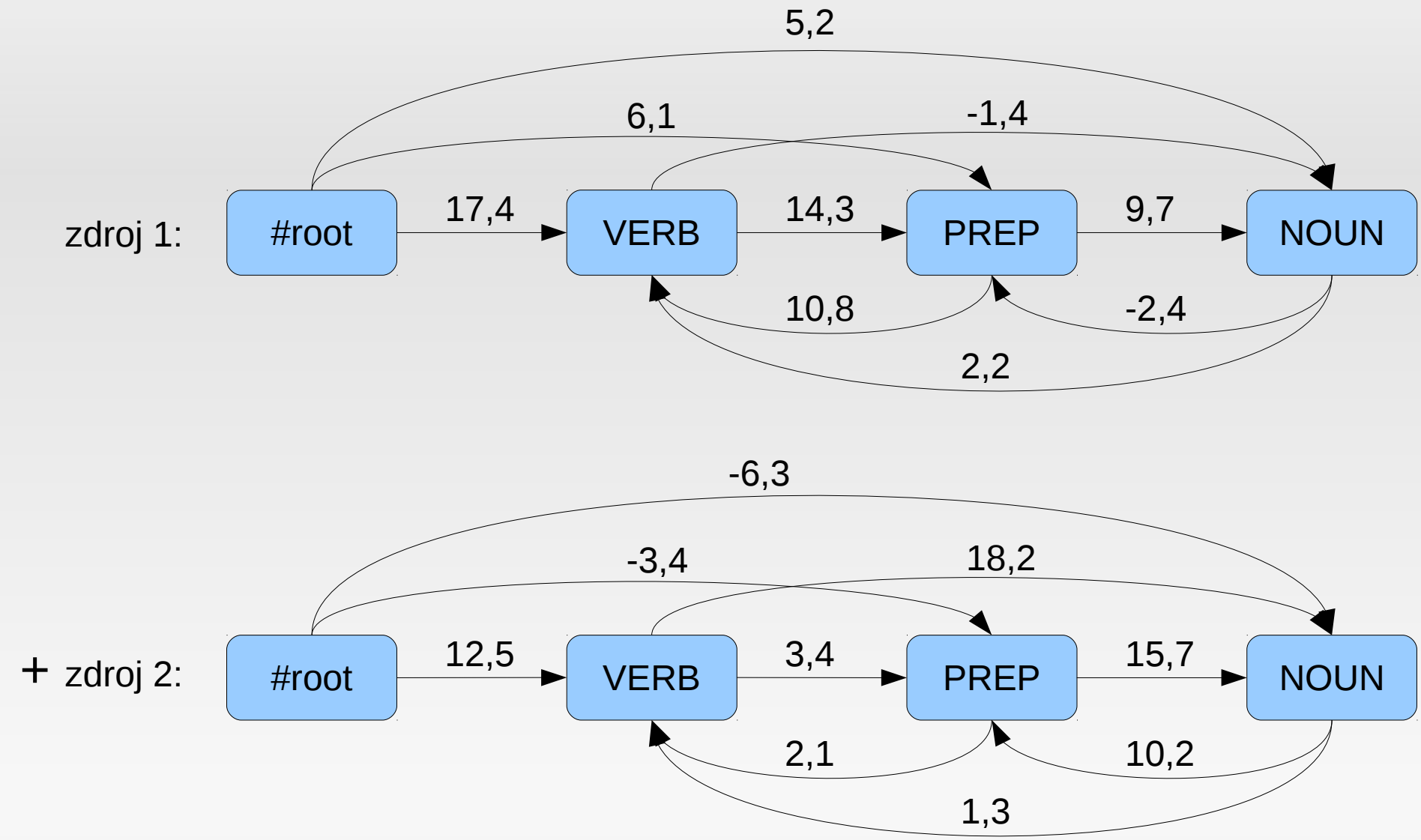
Více zdrojů: kombinace stromů



Více zdrojů: kombinace stromů

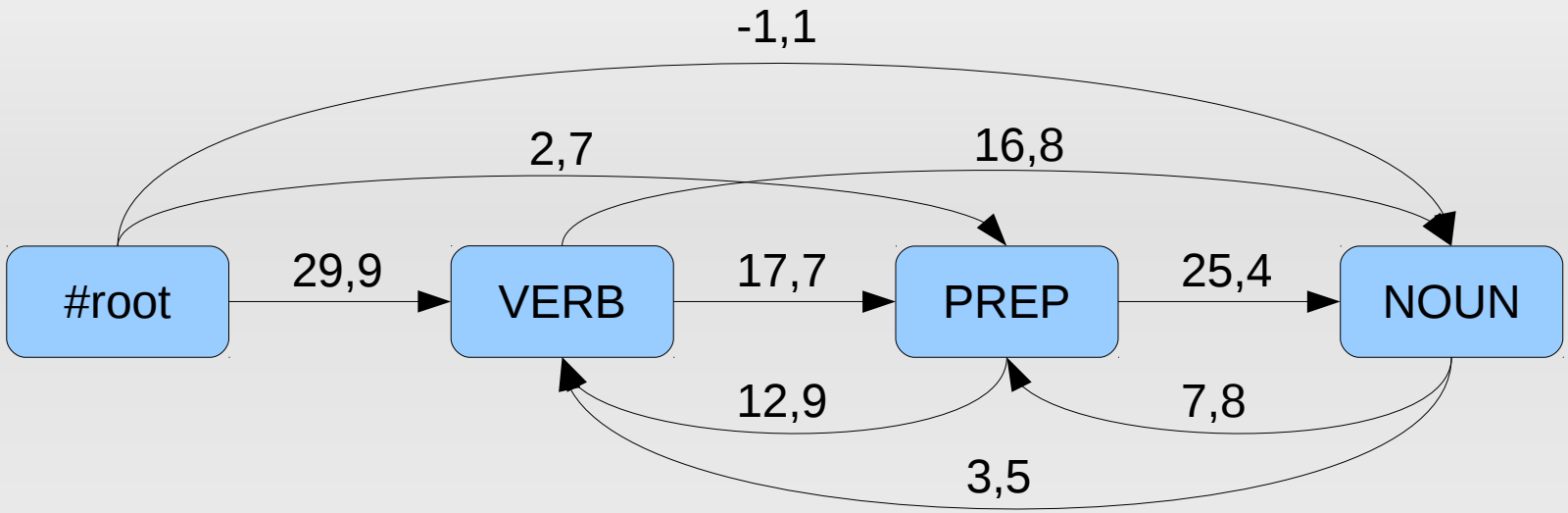


Více zdrojů: interpolace modelů



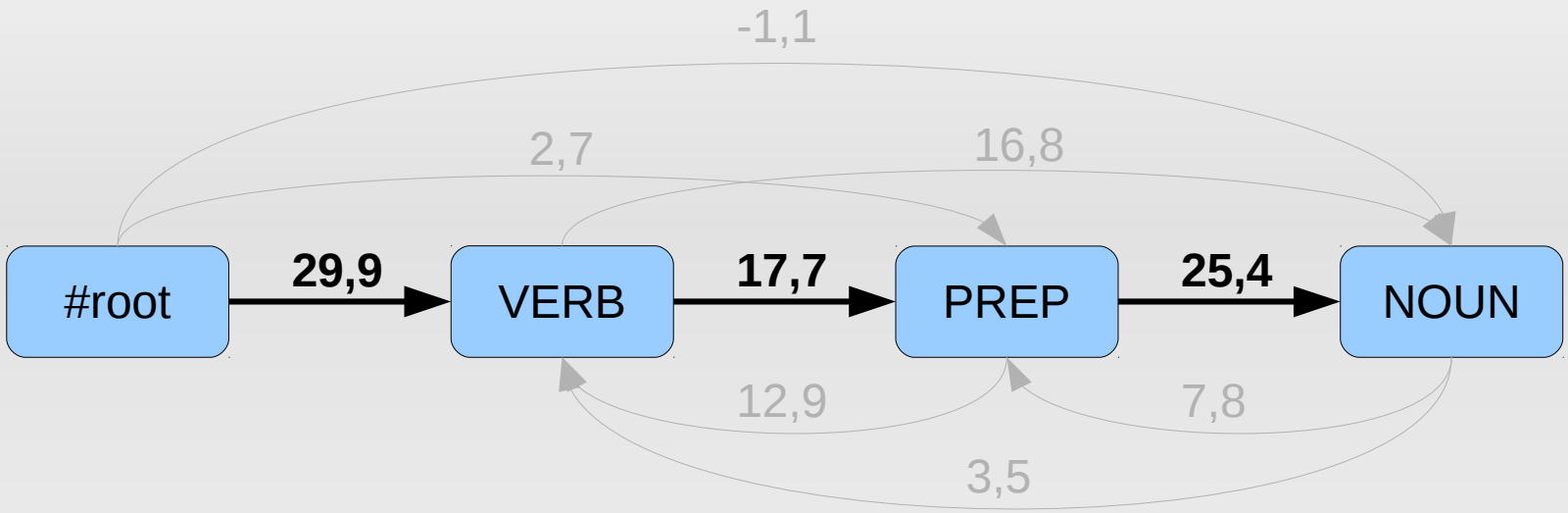
Více zdrojů: interpolace modelů

= **cíl (Σ):**

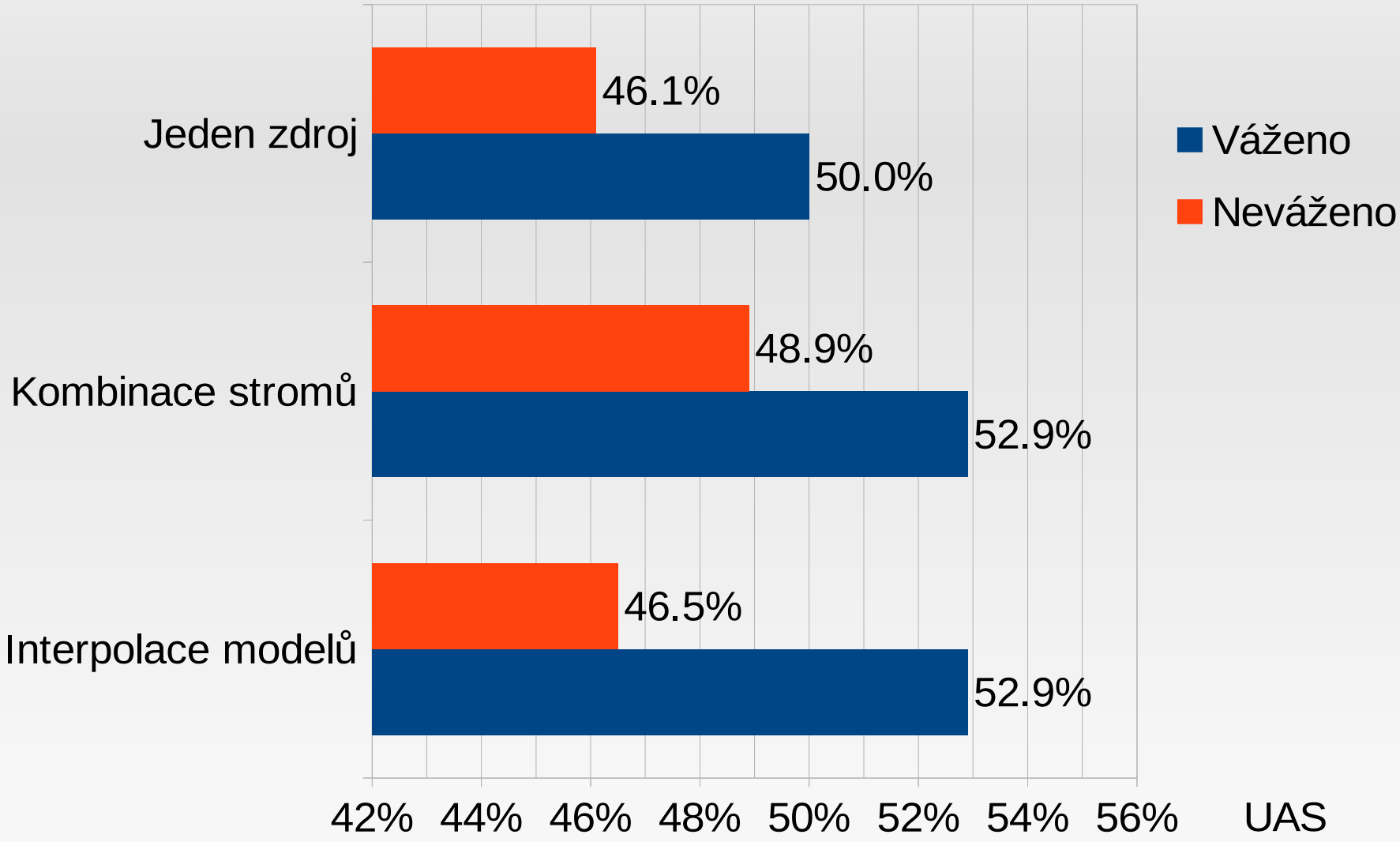


Více zdrojů: interpolace modelů

= cíl:



Porovnání metod

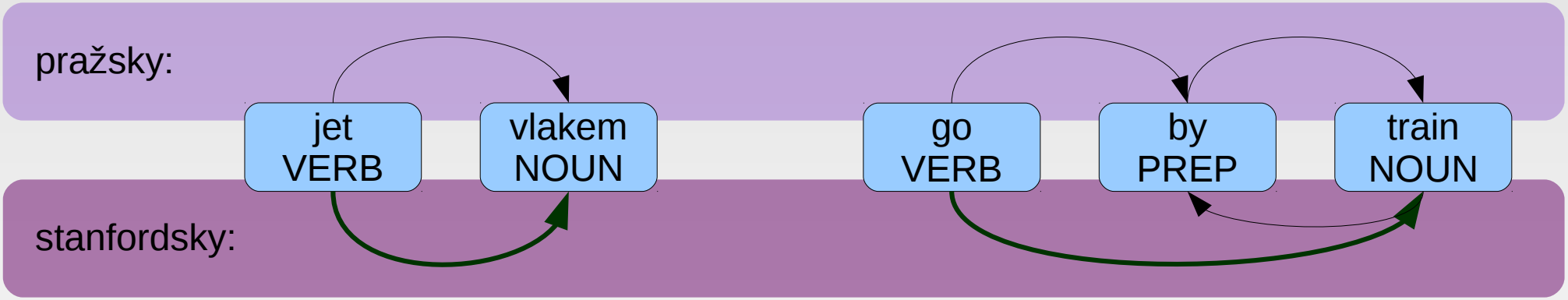


Anotační styl (pro vícezdr. přenos)

- pražský 57% UAS, stanfordský 49% UAS

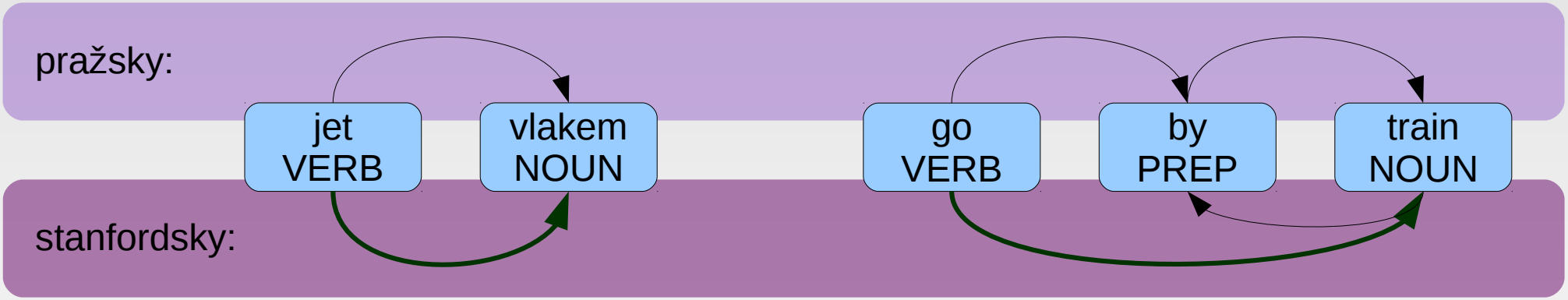
Anotační styl (pro vícezdr. přenos)

- pražský 57% UAS, stanfordský 49% UAS
  - pražský lepší jako základ; předložky stanfordsky?



Anotační styl (pro vícezdr. přenos)

- pražský 57% UAS, stanfordský 49% UAS
  - pražský lepší jako základ; předložky stanfordsky?



- zajímavé výsledky při kombinaci obou stylů anotace předložek (+0,39% UAS)
  - osamocené jazyky: velké zlepšení (et +3%, fa +2%)

Lexikalizace (výhled do budoucna)

- lexikální rysy (forma, lemma) jsou klíčové pro vysokou úspěšnost parseru (cca +6% UAS)
- bez použití paralelních dat
  - self-training
- s použitím paralelních dat/slovníku/překladače
  - parsing paralelní věty, projekce výsledného stromu
  - ukotvení slov (word embeddings, continuous vector space representations)

Závěr

- Přenos delexikalizovaného parseru
  - jeden zdroj
  - více zdrojů: kombinace stromů, interpolace modelů
  - KL_{cpos3}: podobnost jazyků pro výběr/vážení zdrojů
- Anotační styl pro parsing
  - pražský styl celkově lepší než stanfordský
  - stanfordské předložky dobré pro mezijazyčný přenos
- Lexikalizace (v budoucnosti)

Děkuji za pozornost

Rudolf Rosa
rosa@ufal.mff.cuni.cz

**Syntaktická analýza
vět přirozeného jazyka
pomocí částečně řízených metod**

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

http://ufal.mff.cuni.cz/rudolf-rosa/