Rudolf Rosa
rosa@ufal.mff.cuni.cz

Parsing Natural Language Sentences by Semi-supervised Methods

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Dissertation interim report defence, Praha, 15 June 2015

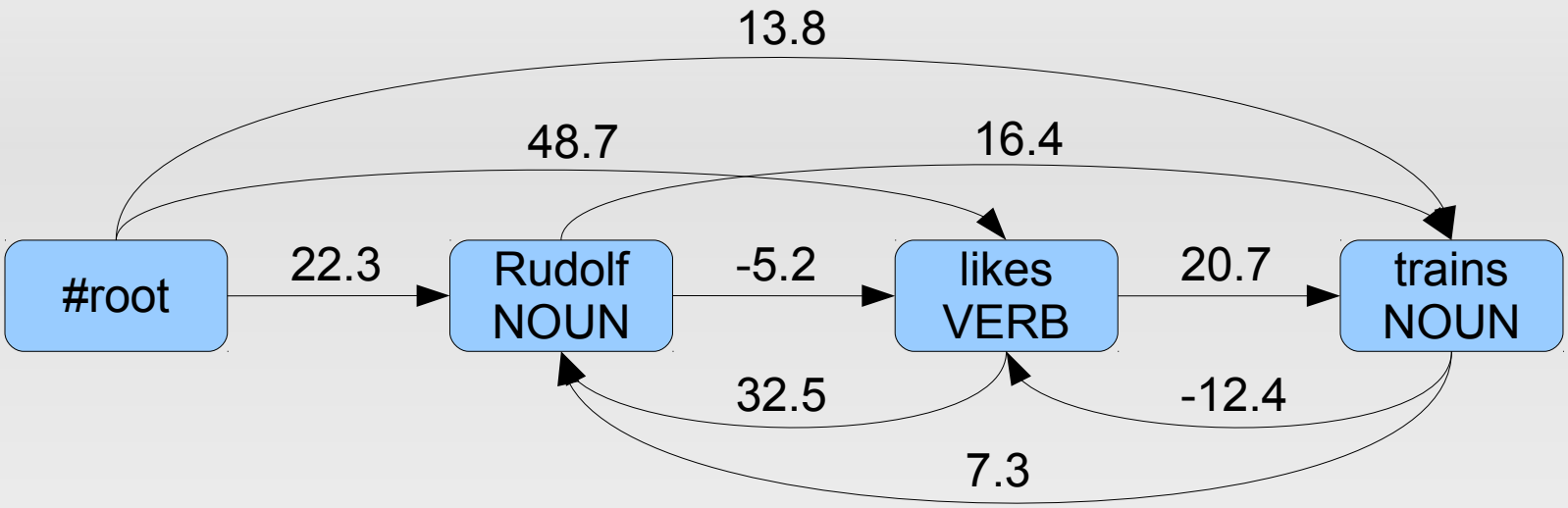
Outline

- Intro and motivation
- MSTParser and its delexicalization
- Delexicalized parser transfer
  - single-source
  - multi-source
    - tree combination
    - model interpolation
- Choice of parsing annotation style

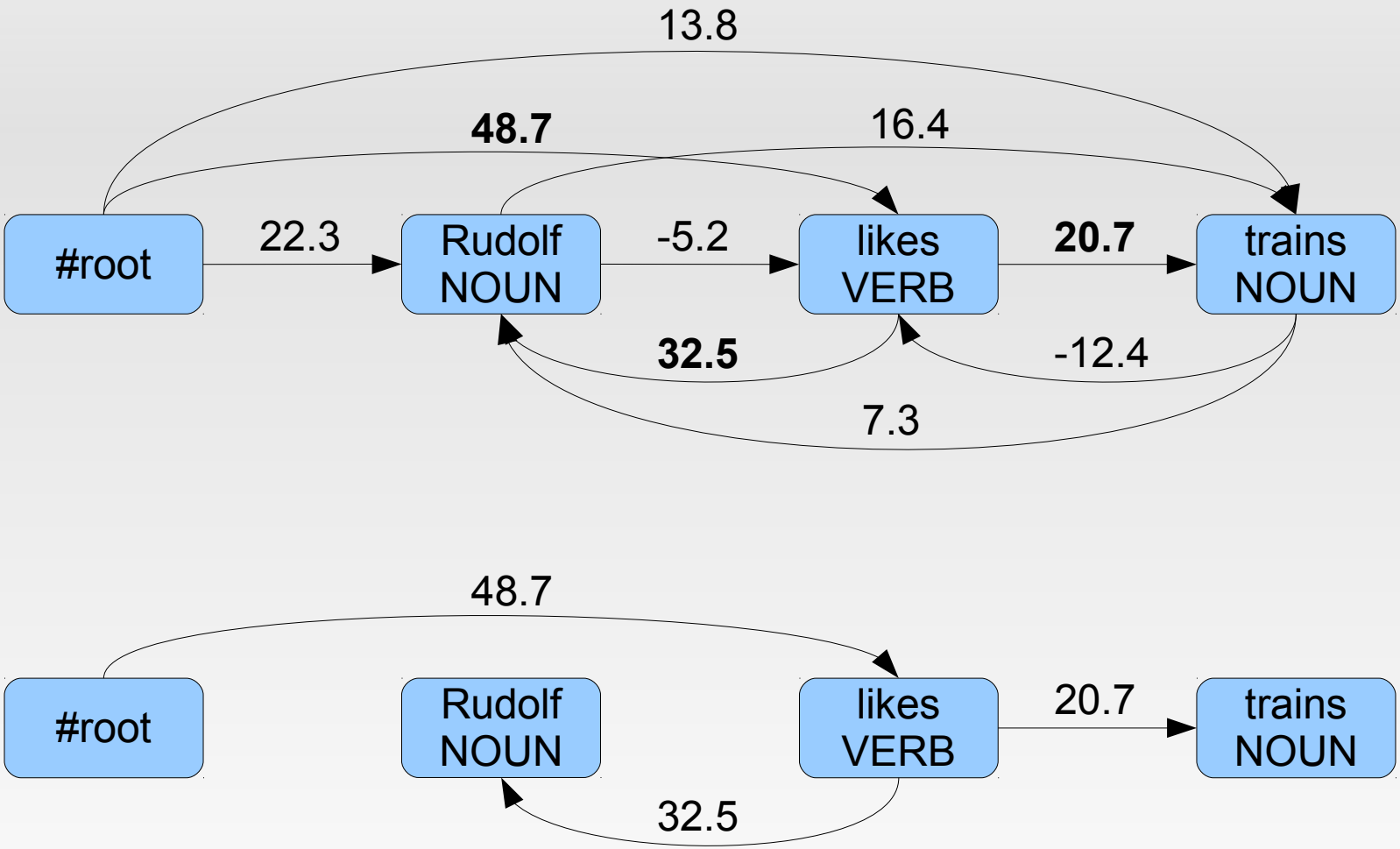
Semi-supervised parsing

- fully supervised dependency parsing
  - requires training data (treebank) or a grammar
  - there are ~100 treebanks (manually annotated)
  - there are ~7 000 languages
  - + various domains, language evolution...
- semi-supervised parsing
  - utilize existing resources, avoid new annotations
    - treebanks for other langs (HamleDT: 30 langs)
    - unannotated data (here: POS tagged)

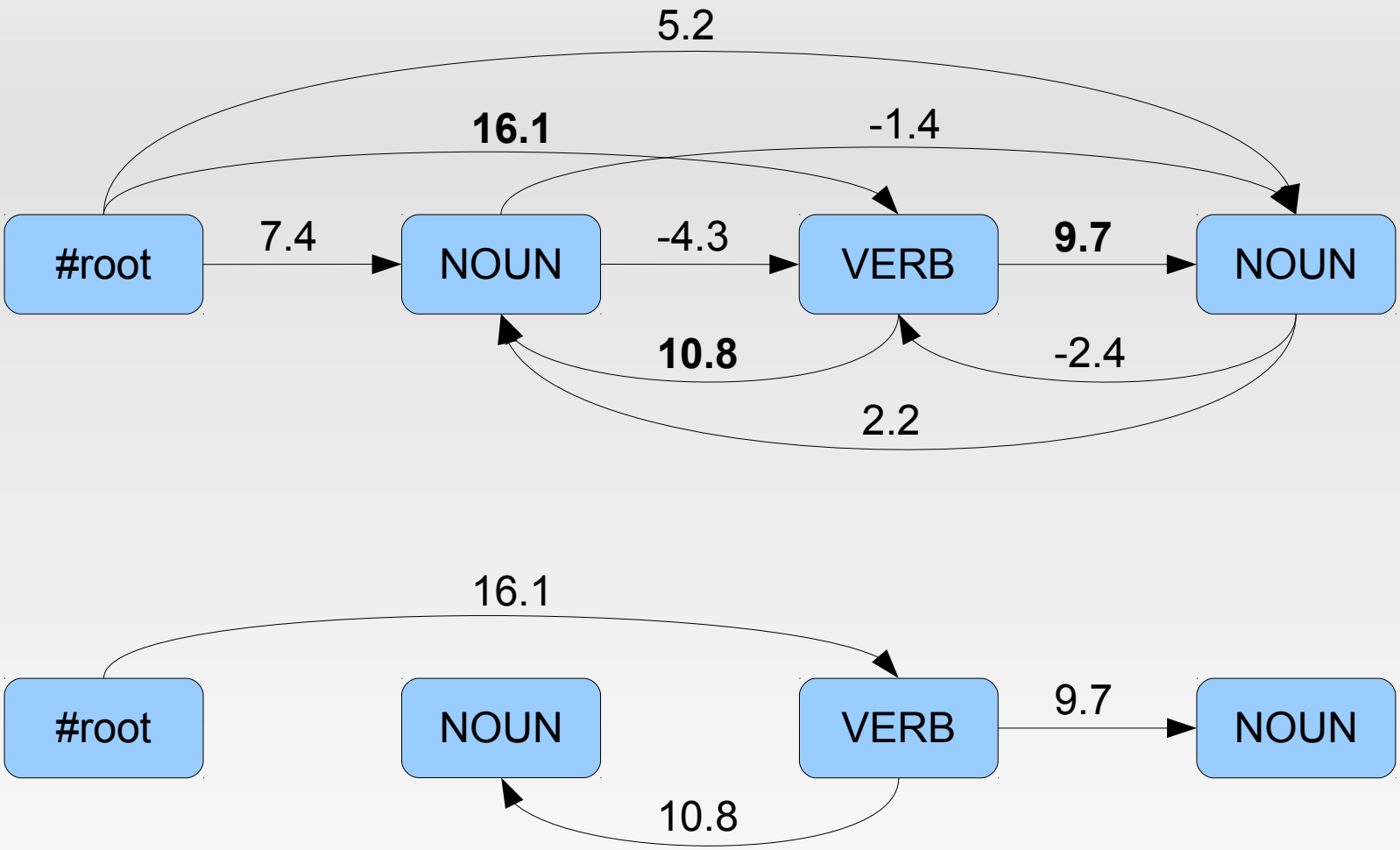
(Lexicalized) MSTParser



(Lexicalized) MSTParser



Delexicalized MSTParser



Single-source delex parser transfer

- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)

Single-source delex parser transfer

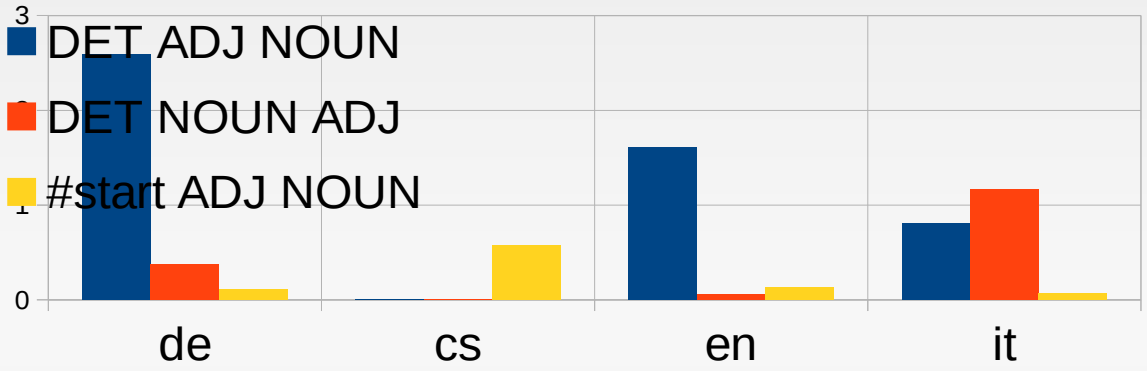
- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)
- how to choose the src lang for a tgt lang?

Single-source delex parser transfer

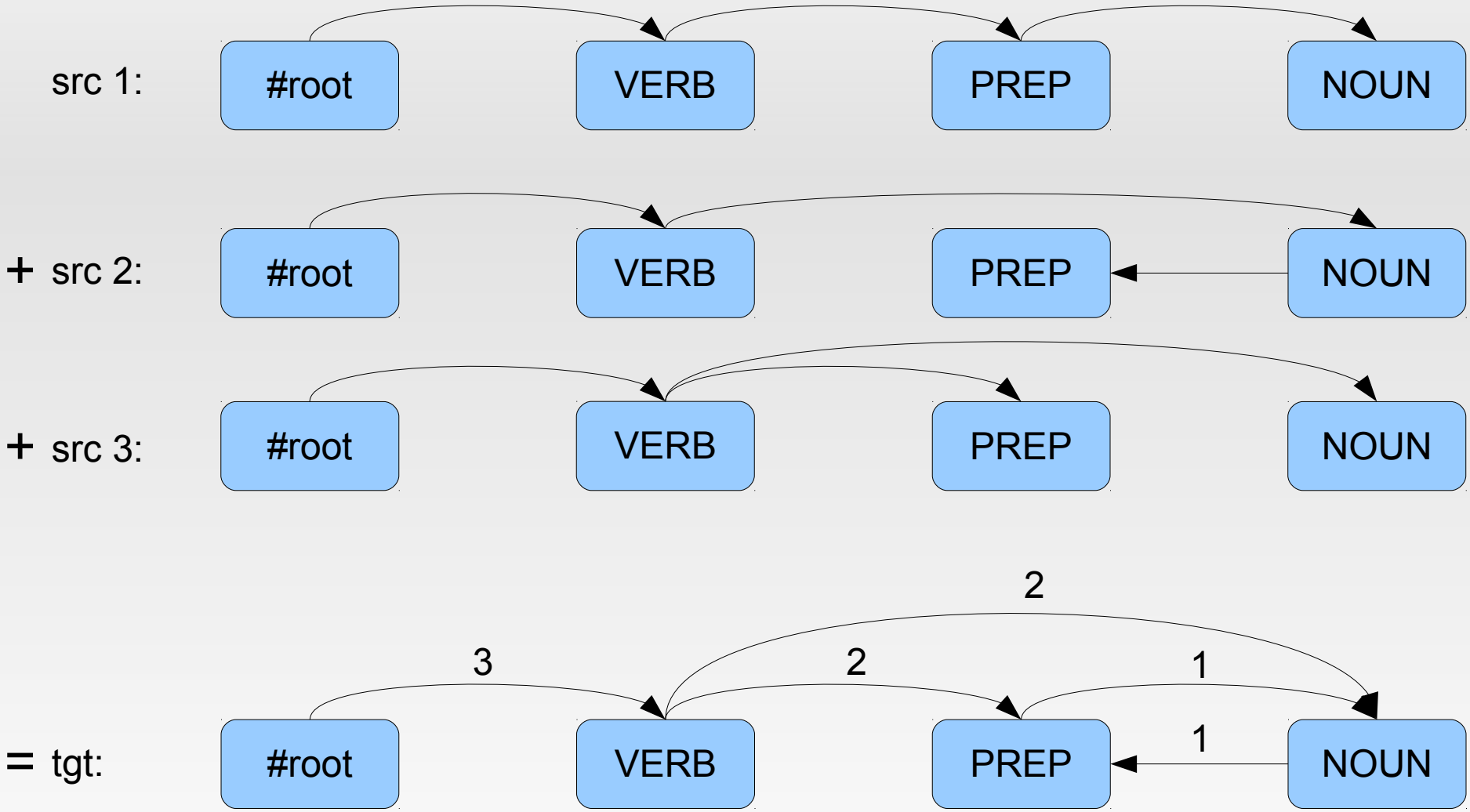
- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)
- how to choose the src lang for a tgt lang?
  - src should be as similar to tgt as possible

Single-source delex parser transfer

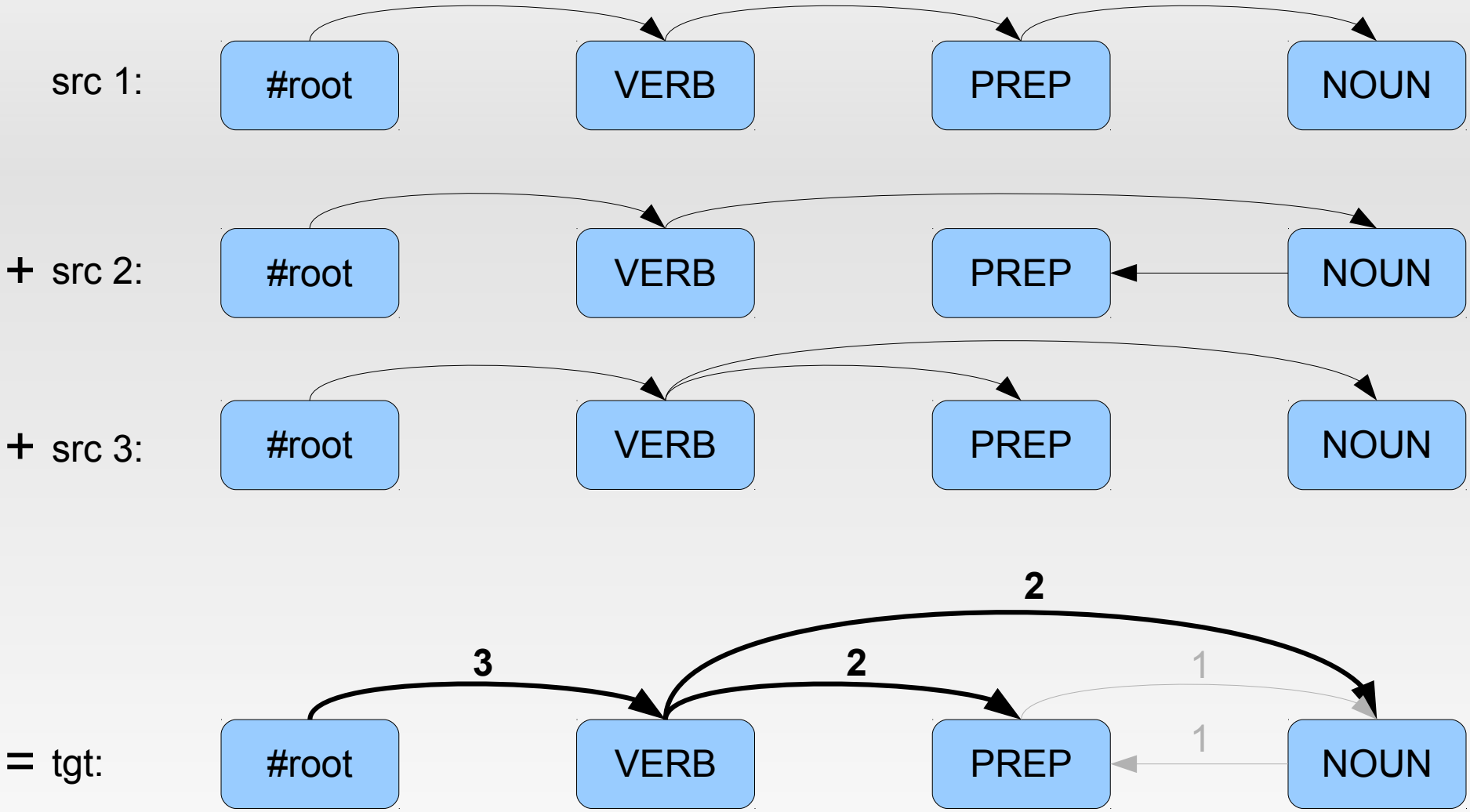
- train a delex parser on a src lang treebank
- apply to a tgt lang (-treebank, +POS tagger)
- how to choose the src lang for a tgt lang?
  - src should be as similar to tgt as possible
  - KL_{cpos3}: POS trigram distribution in tagged corpora



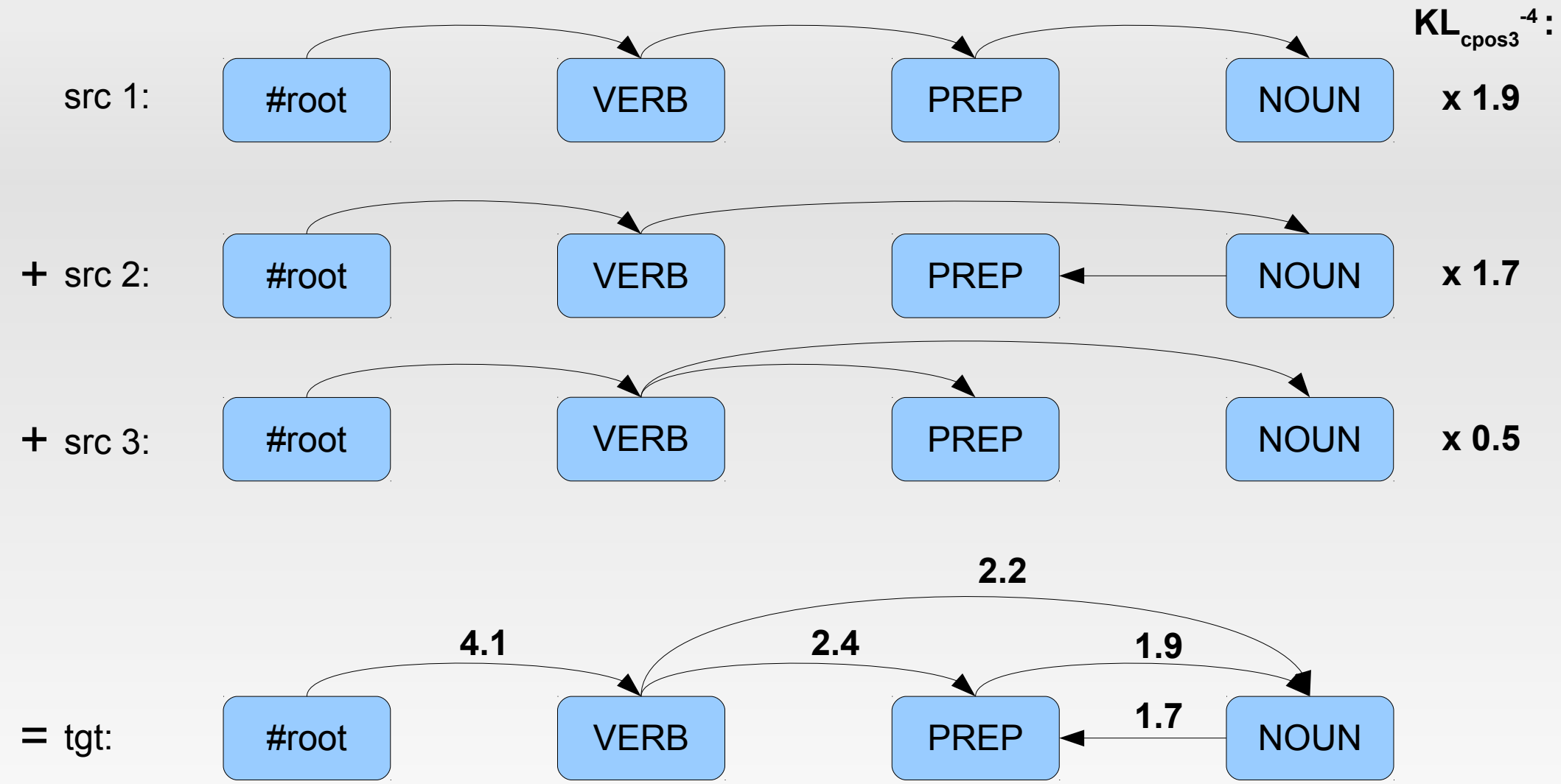
Multi-source: tree combination



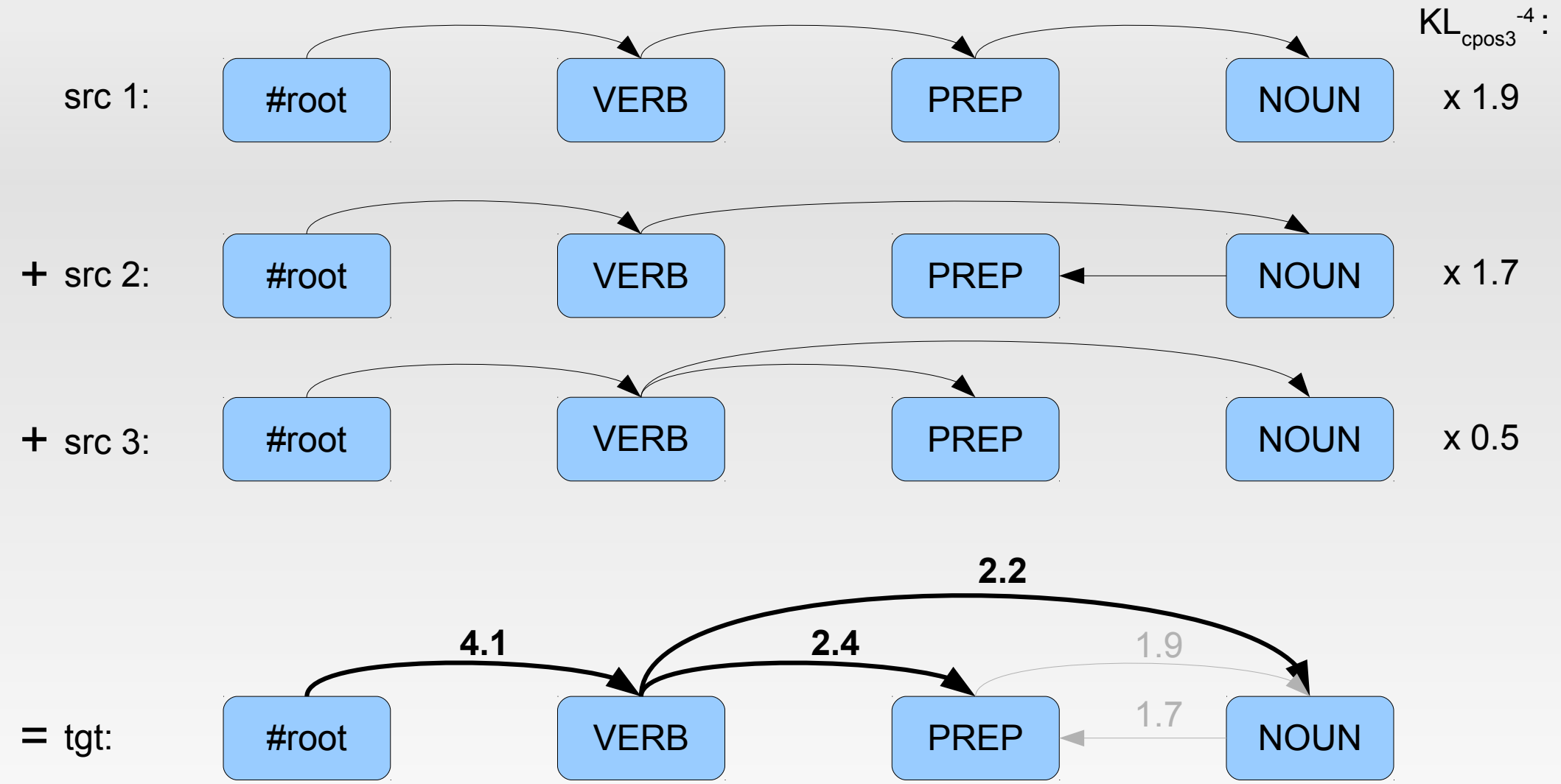
Multi-source: tree combination



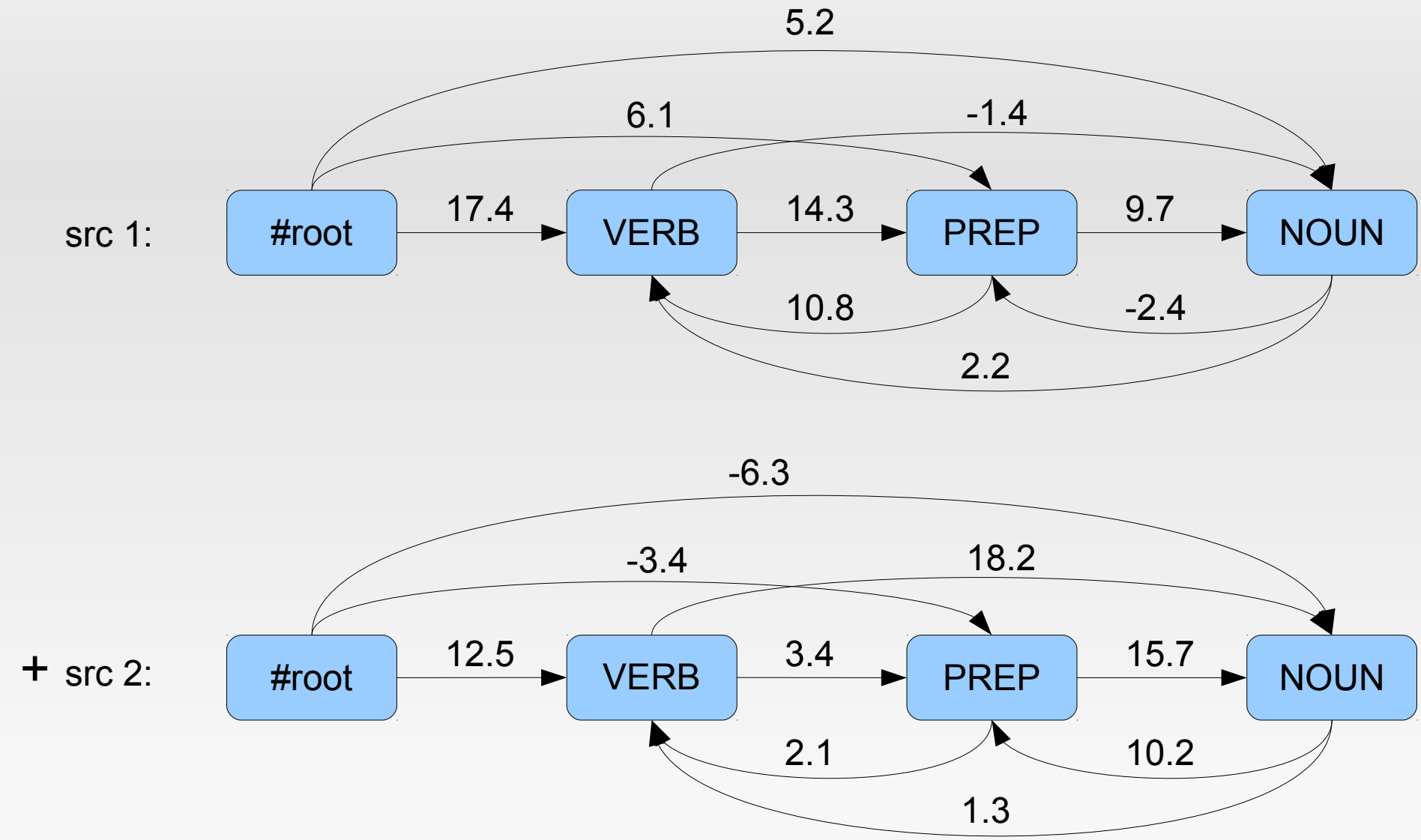
Multi-source: tree combination



Multi-source: tree combination

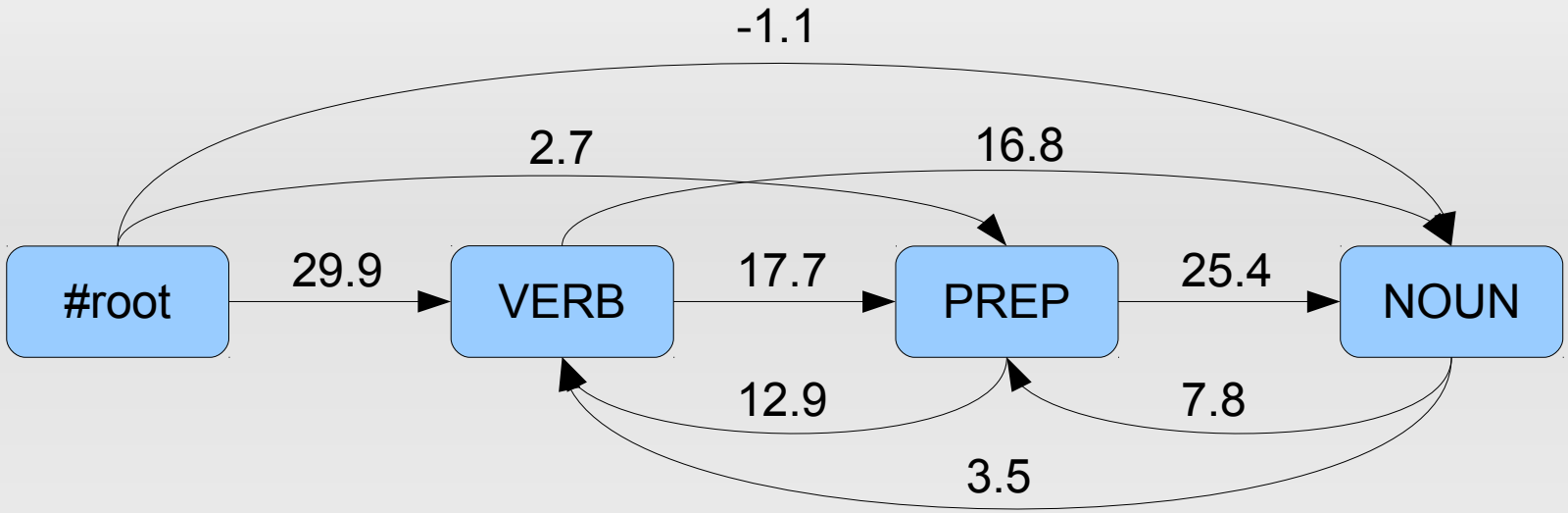


Multi-source: model interpolation



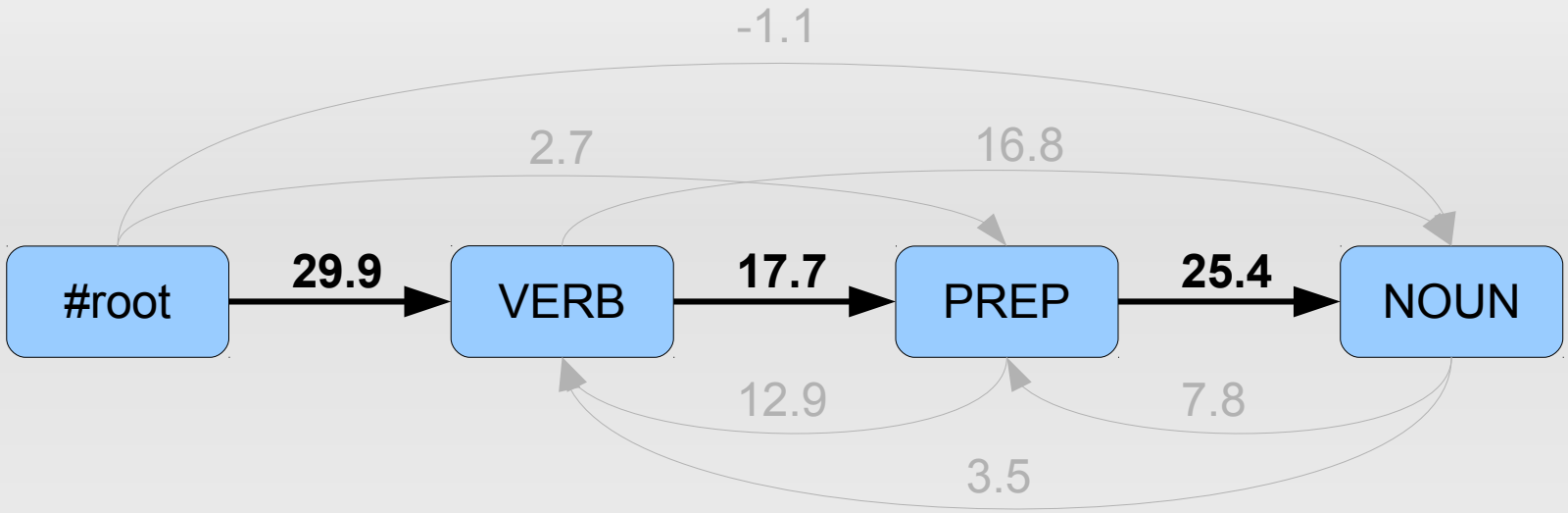
Multi-source: model interpolation

= **tgt (∑):**

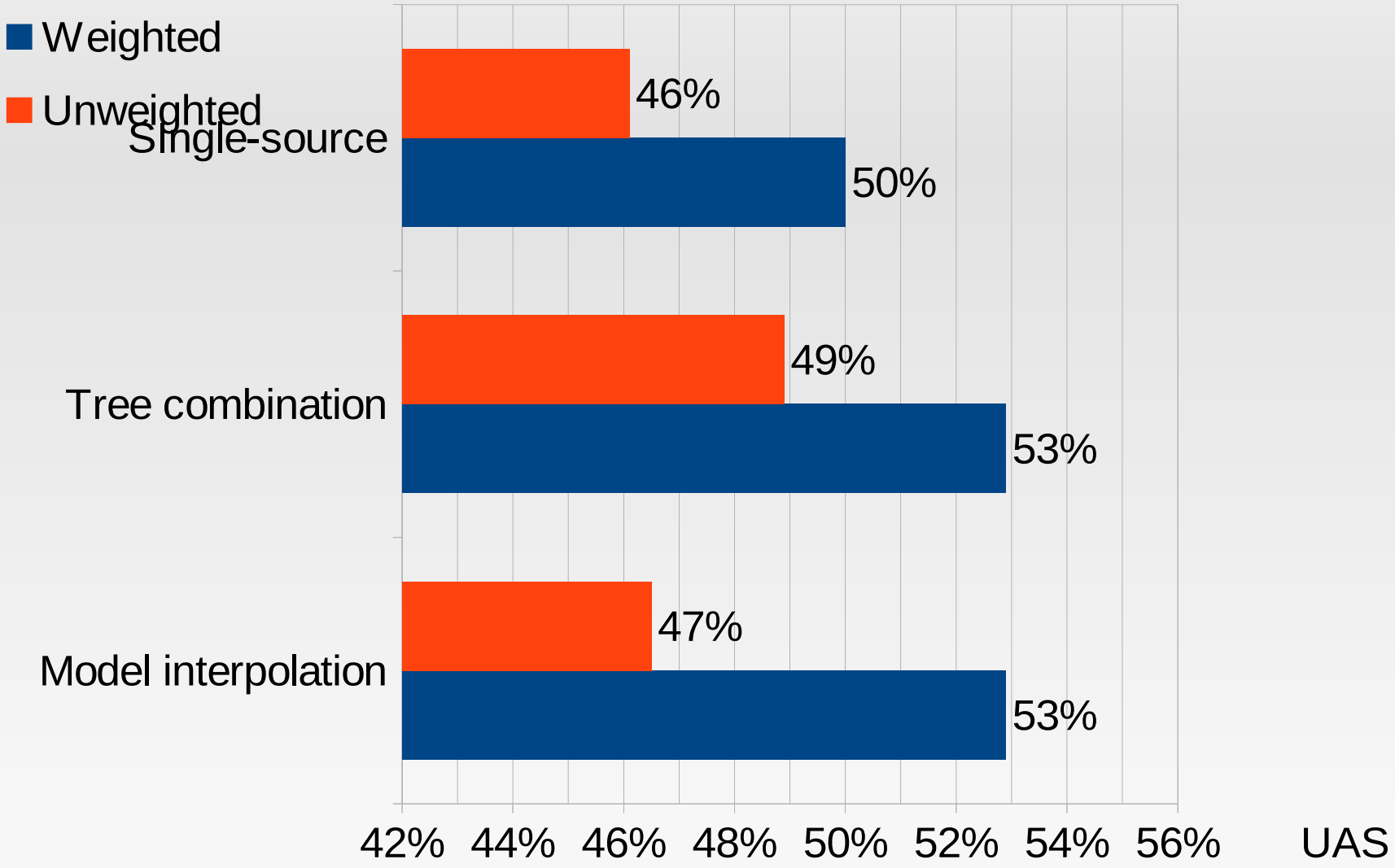


Multi-source: model interpolation

= tgt:



Evaluation on HamleDT (30 langs)

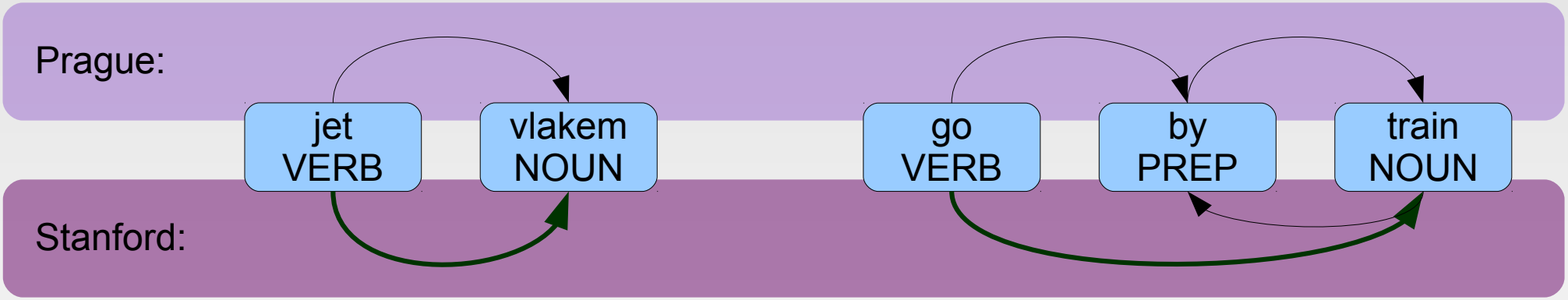


Annotation style (multi-source)

- Prague 57% UAS, Stanford 49% UAS

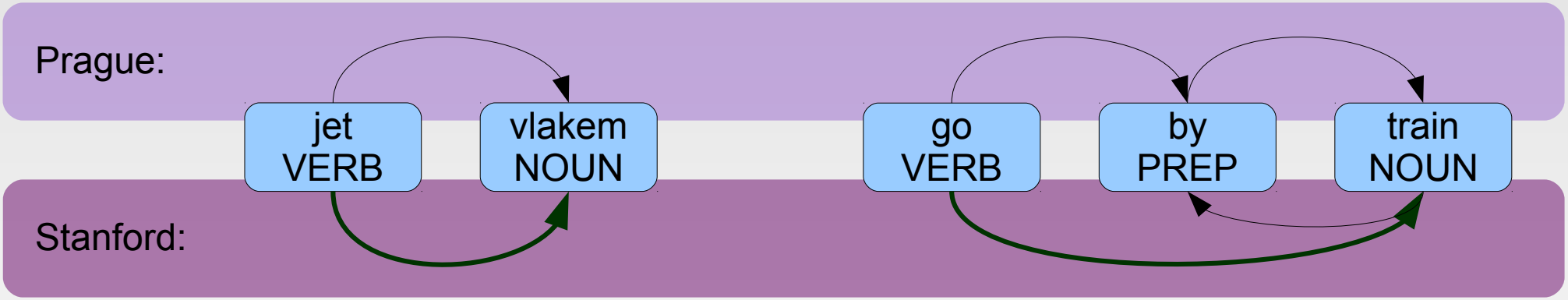
Annotation style (multi-source)

- Prague 57% UAS, Stanford 49% UAS
  - Prague better as base; Stanford adpositions?



Annotation style (multi-source)

- Prague 57% UAS, Stanford 49% UAS
  - Prague better as base; Stanford adpositions?



- interesting results when combining both adposition annotation styles (+0.39% UAS)
  - solitary langs: big improvement (et +3%, fa +2%)

Conclusion

- Delexicalized parser transfer
  - single-source
  - multi-source: tree combination, model interpolation
  - KL_{cpos3}: lang similarity for src selection/weighting
- Annotation style for parsing
  - Prague better than Stanford
  - Stanford adpositions good for cross-lingual transfer

Thank you for your attention

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Parsing Natural Language Sentences by Semi-supervised Methods

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

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