Rudolf Rosa, Zdeněk Žabokrtský
{rosa,zabokrtsky}@ufal.mff.cuni.cz

KL_{cpos^3}

a Language Similarity Measure for Delexicalized Parser Transfer

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

ACL, Beijing, 28 July 2015

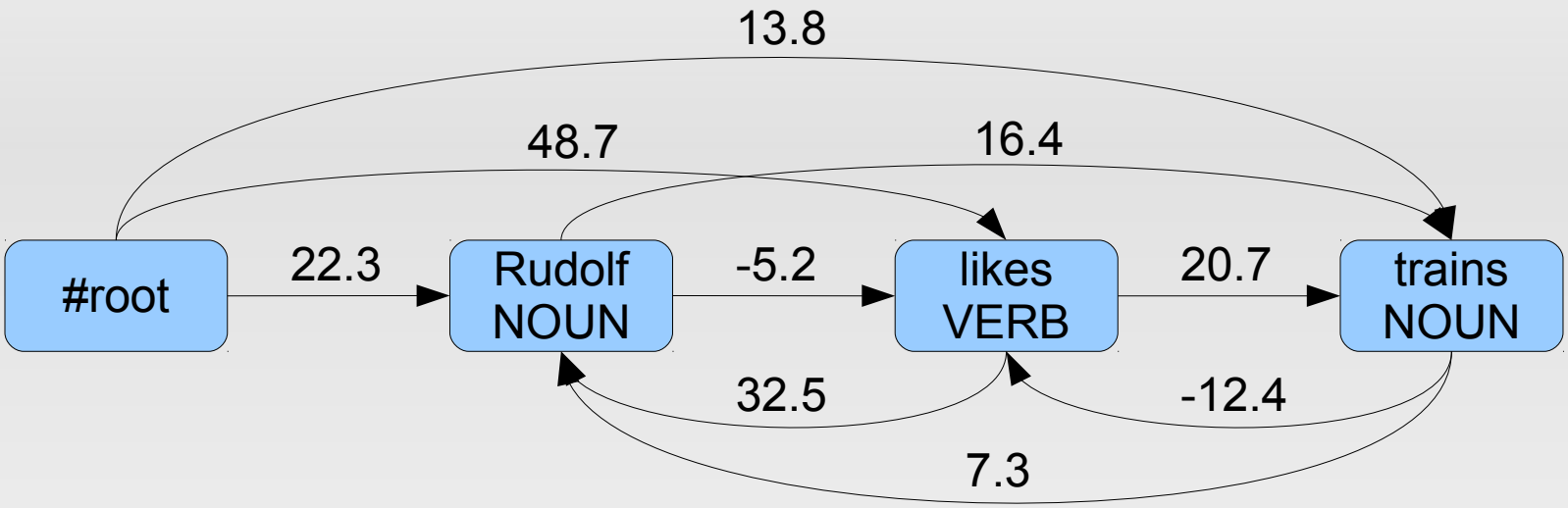
Outline

- Intro and motivation
- MSTParser and its delexicalization
- Single-source delexicalized parser transfer
  - $KL_{cpos\ 3}$ for source selection
- Multi-source delexicalized parser transfer
  - $KL_{cpos\ 3}$ for source weighting
- Results

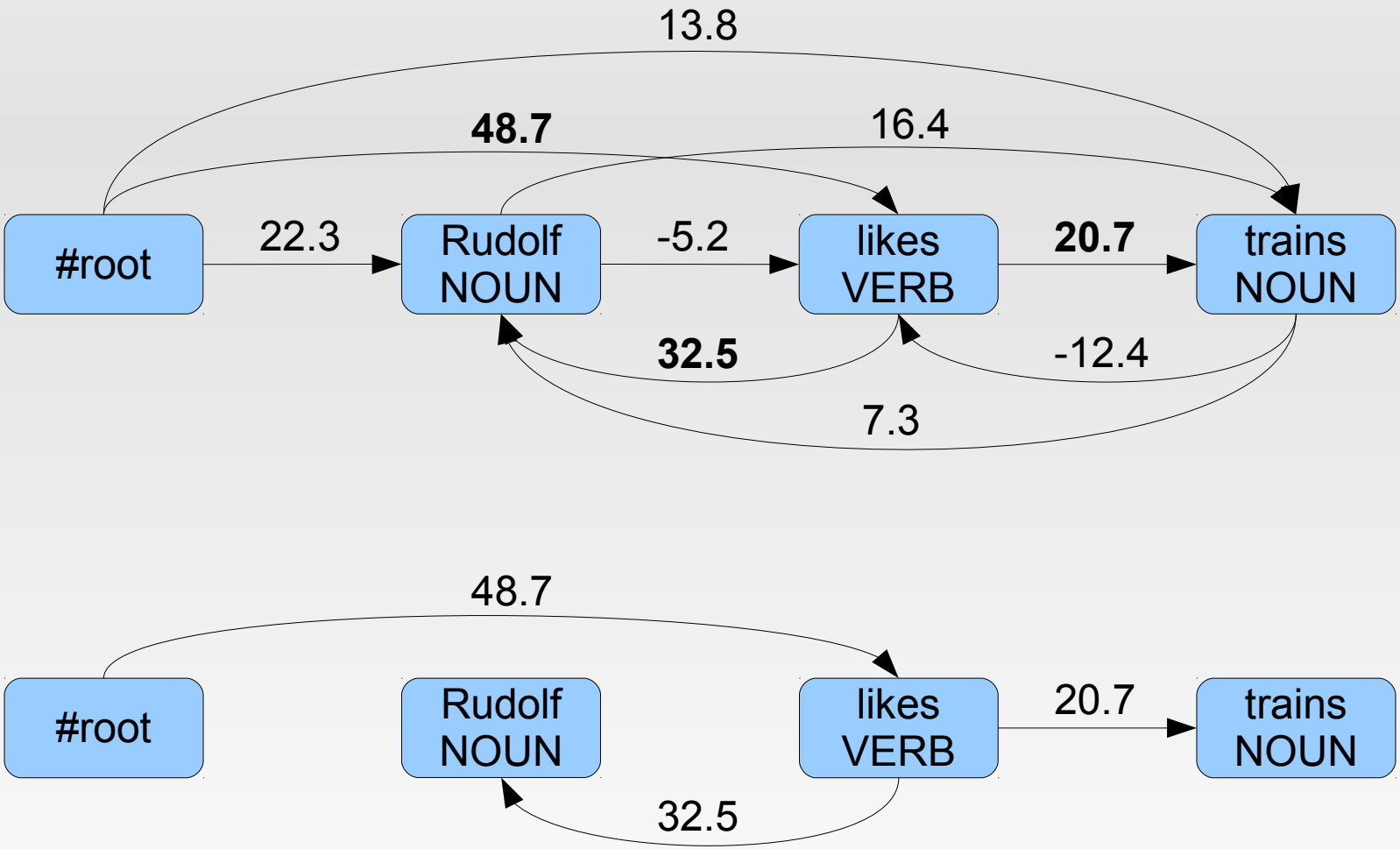
Semi-supervised parsing

- 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)

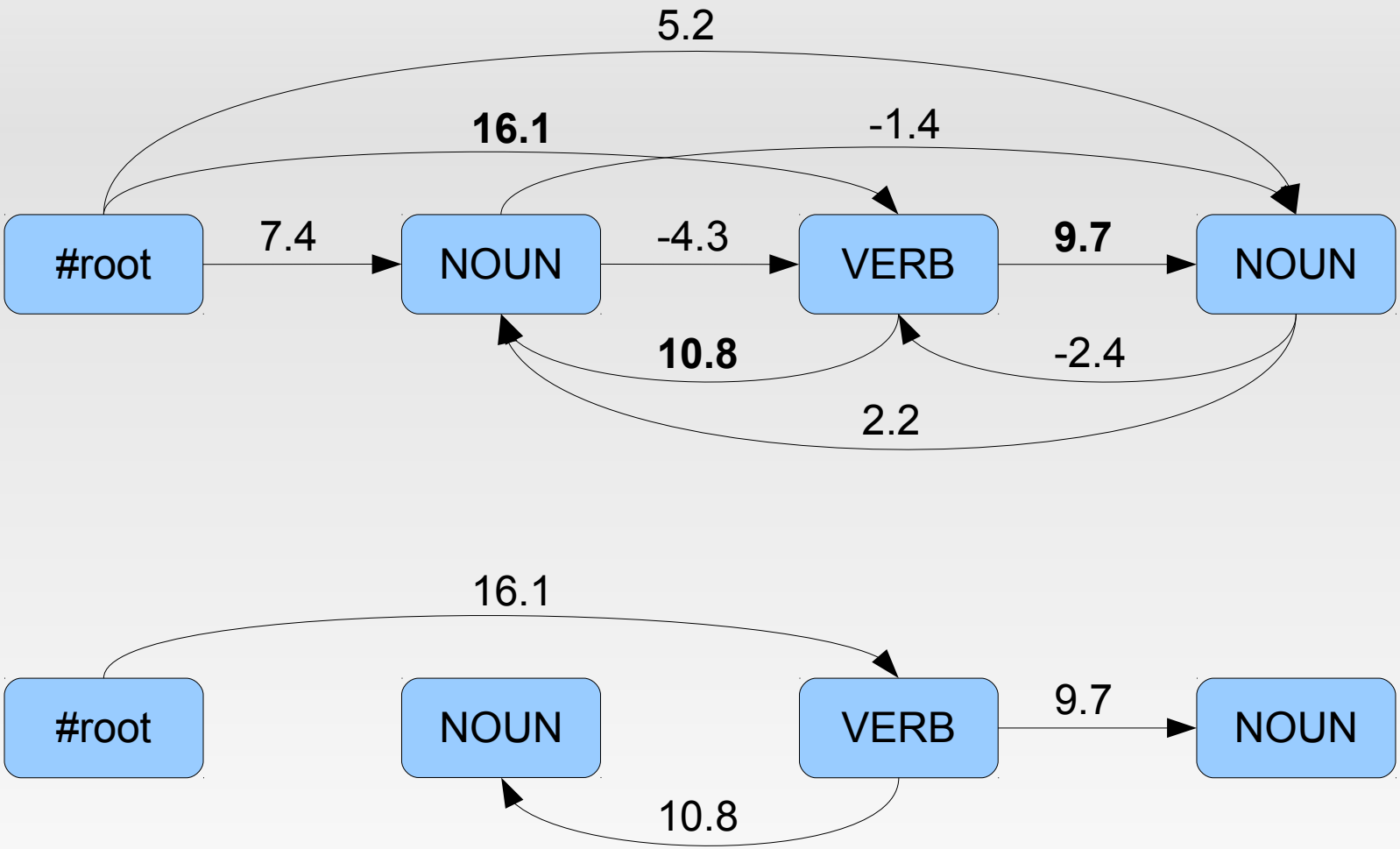
(Lexicalized) MSTParser



(Lexicalized) MSTParser



Delexicalized MSTParser



Single-source delex parser transfer

- (Zeman and Resnik, 2008)
- train a delexicalized parser on a **source** language treebank (e.g. Czech – PDT)
- apply it to a **target** language, without a treebank but with a POS tagger (e.g. Slovak)

Utilizing multiple treebanks

- HamleDT: 30 harmonized treebanks
  - (split: 12 development TBs, 18 testing TBs)
- How do we choose the source treebank?
- Can we use more/all source treebanks?

Utilizing multiple treebanks

- HamleDT: 30 harmonized treebanks
  - (split: 12 development TBs, 18 testing TBs)
- How do we choose the source treebank?
- Can we use more/all source treebanks?

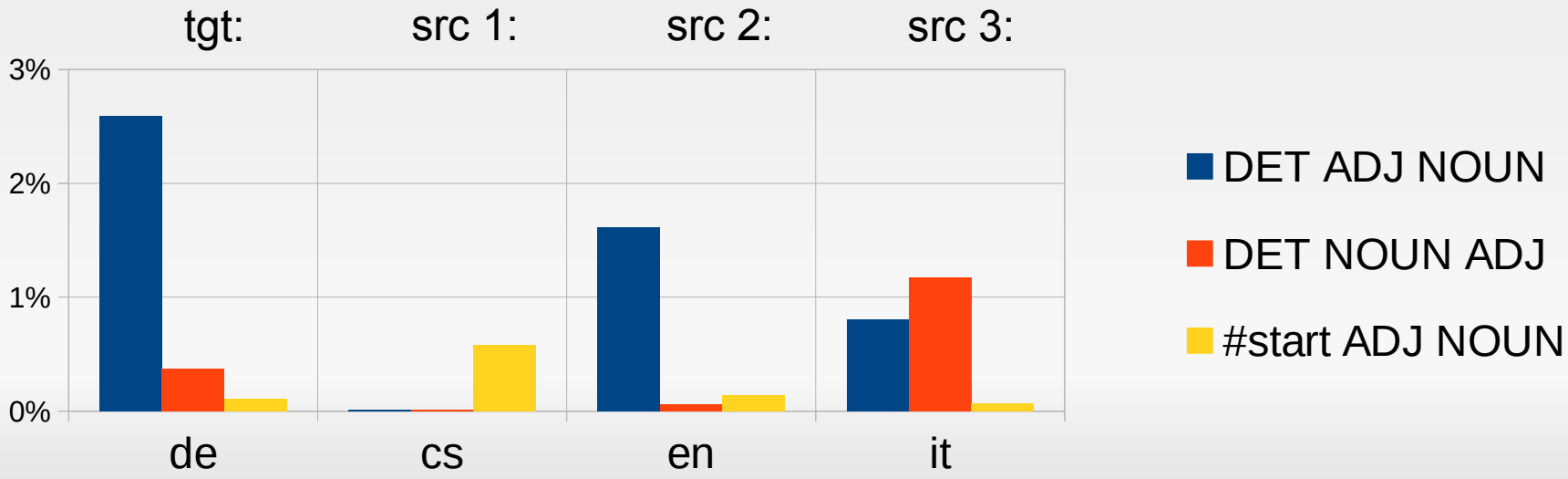
- treebank concatenation (McDonald et al., 2011)
  - if you don't know what to do,
    just concatenate all the data you have
  - concatenate all source treebanks → train a parser →
    → apply the parser to the target text

Choosing the source treebank

- src should be as similar to tgt as possible
  - WALS (Naseem et al., 2012)
  - POS *n*-gram model (Søgaard and Wulff, 2012)

Choosing the source treebank

- src should be as similar to tgt as possible
  - WALS (Naseem et al., 2012)
  - POS *n*-gram model (Søgaard and Wulff, 2012)
  - $KL_{cpos\ 3}$(***tgt***, ***src***): Kullback-Leibler divergence of POS trigram distributions



$KL_{cpos\ 3}$

$$KL_{cpos^3}(tgt, src) = \sum_{\forall cpos^3 \in tgt} f_{tgt}(cpos^3) \cdot \log\left(\frac{f_{tgt}(cpos^3)}{f_{src}(cpos^3)}\right)$$

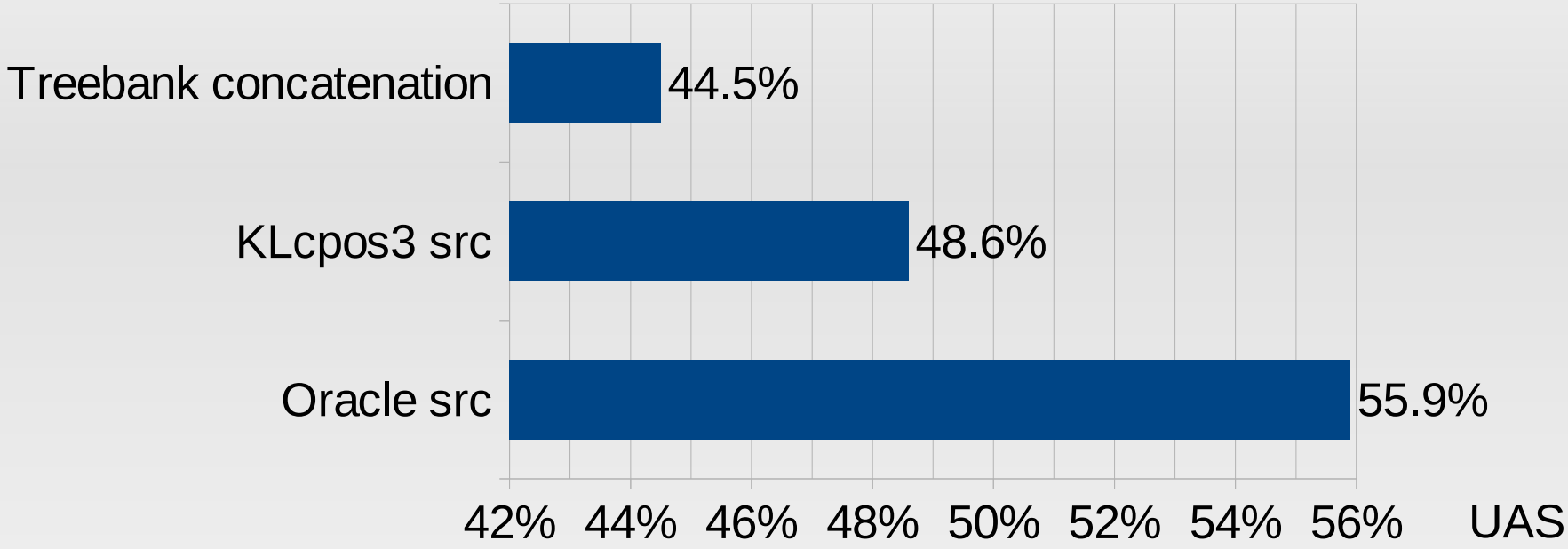
$$cpos^3 = \langle cpos_{i-1}, cpos_i, cpos_{i+1} \rangle$$

$$f(cpos^3) = \frac{count(cpos^3)}{|corpus|}$$

Sample of results (HamleDT)

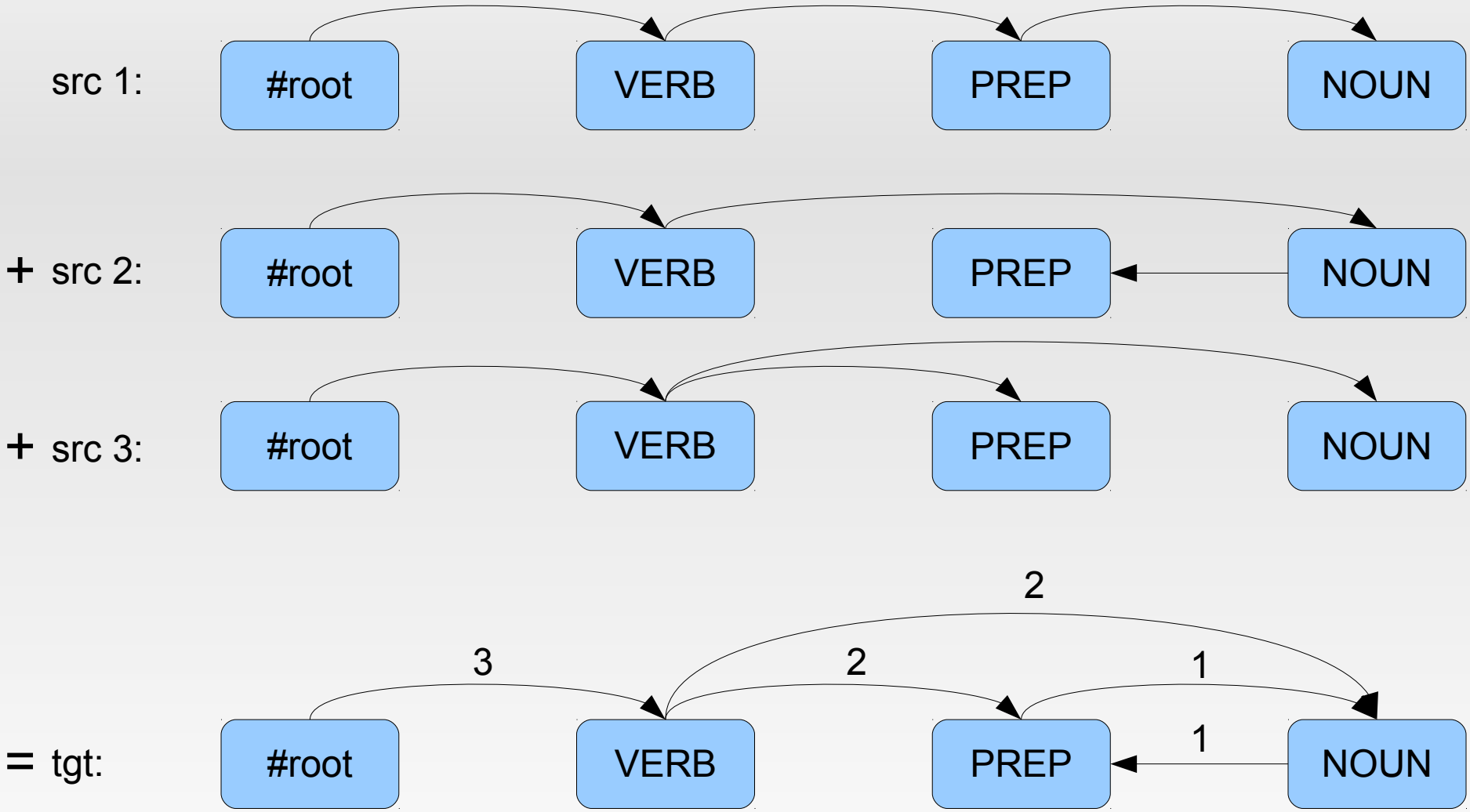
Target lang.	$KL_{cpos\ 3}$ selected src		Oracle (best possible src)
	lang.	UAS	
Bengali	Telugu	66.7	✓
Czech	Slovak	65.8	✓
Danish	Slovenian	42.1	+13.3 English
German	English	56.8	✓
Slovak	Slovenian	58.4	+ 3.3 Czech
Tamil	Turkish	31.1	+22.4 Hindi

Average over 18 test TBs

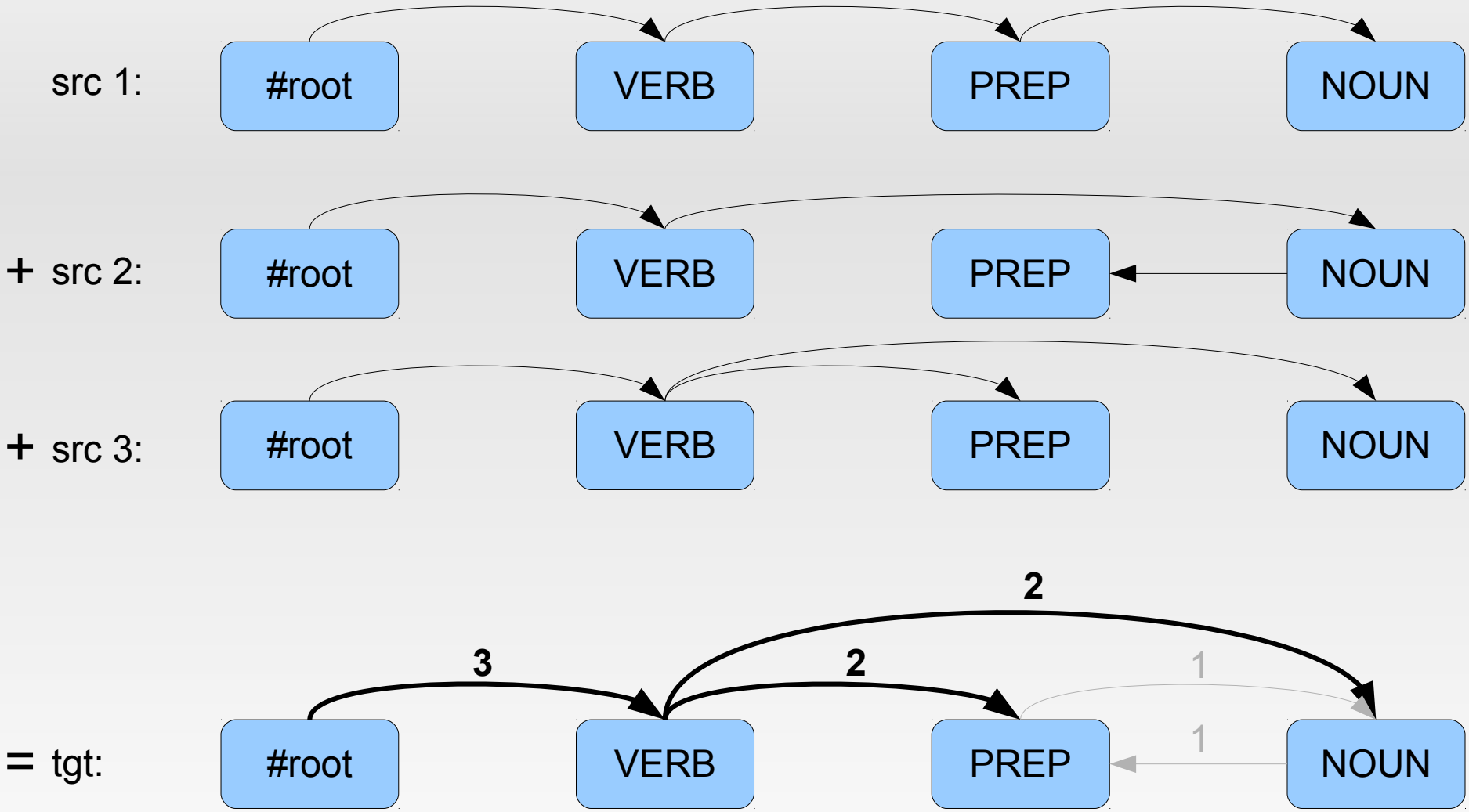


- oracle src identified by $KL_{cpos\ 3}$ in 8 cases
- average best $KL_{cpos\ 3}$ = 0.7
  - < 0.7 → 7x oracle, 2x competitive, 3x bad
  - > 0.7 → 1x oracle, 5x bad

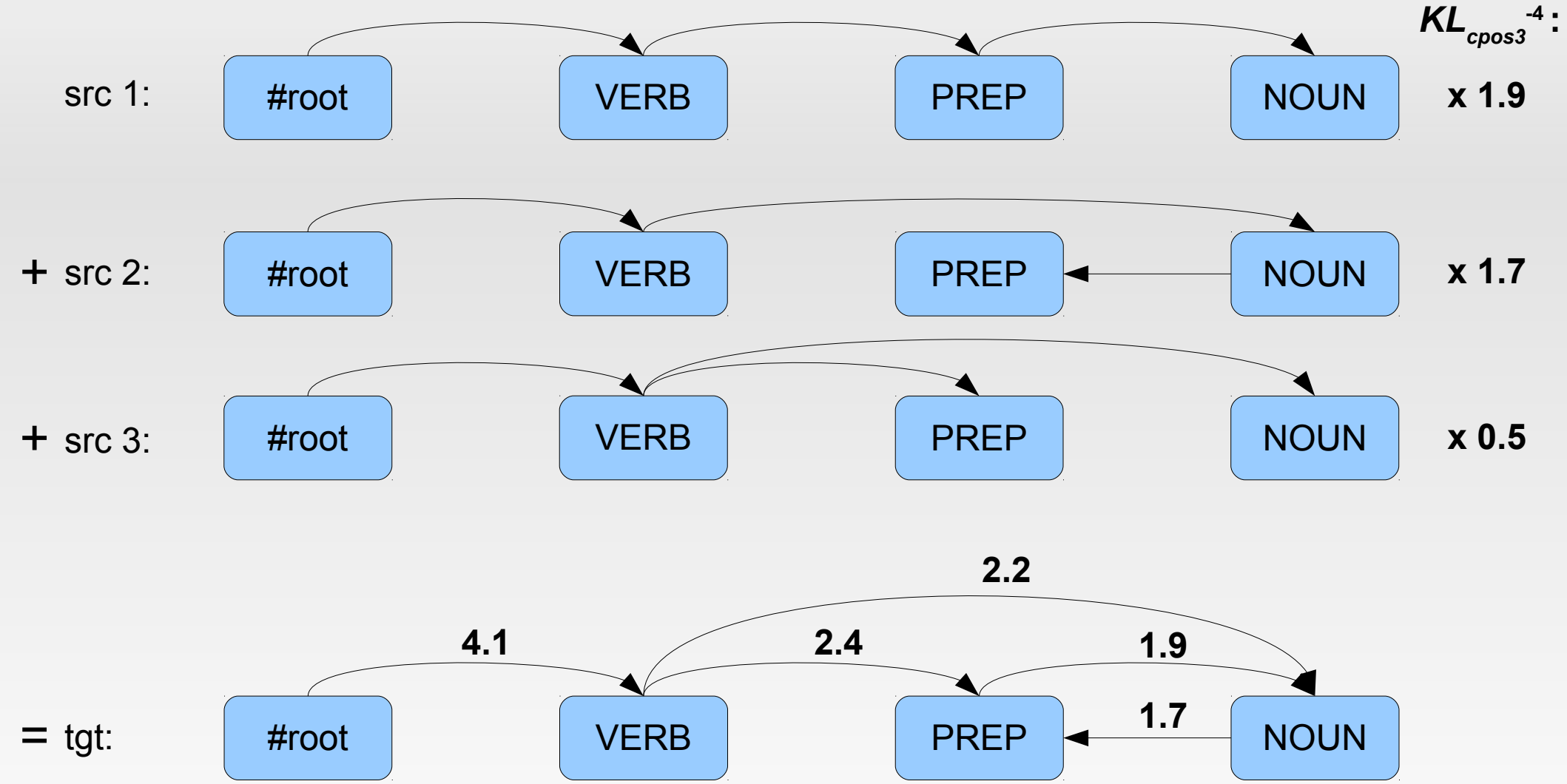
Parse tree combination



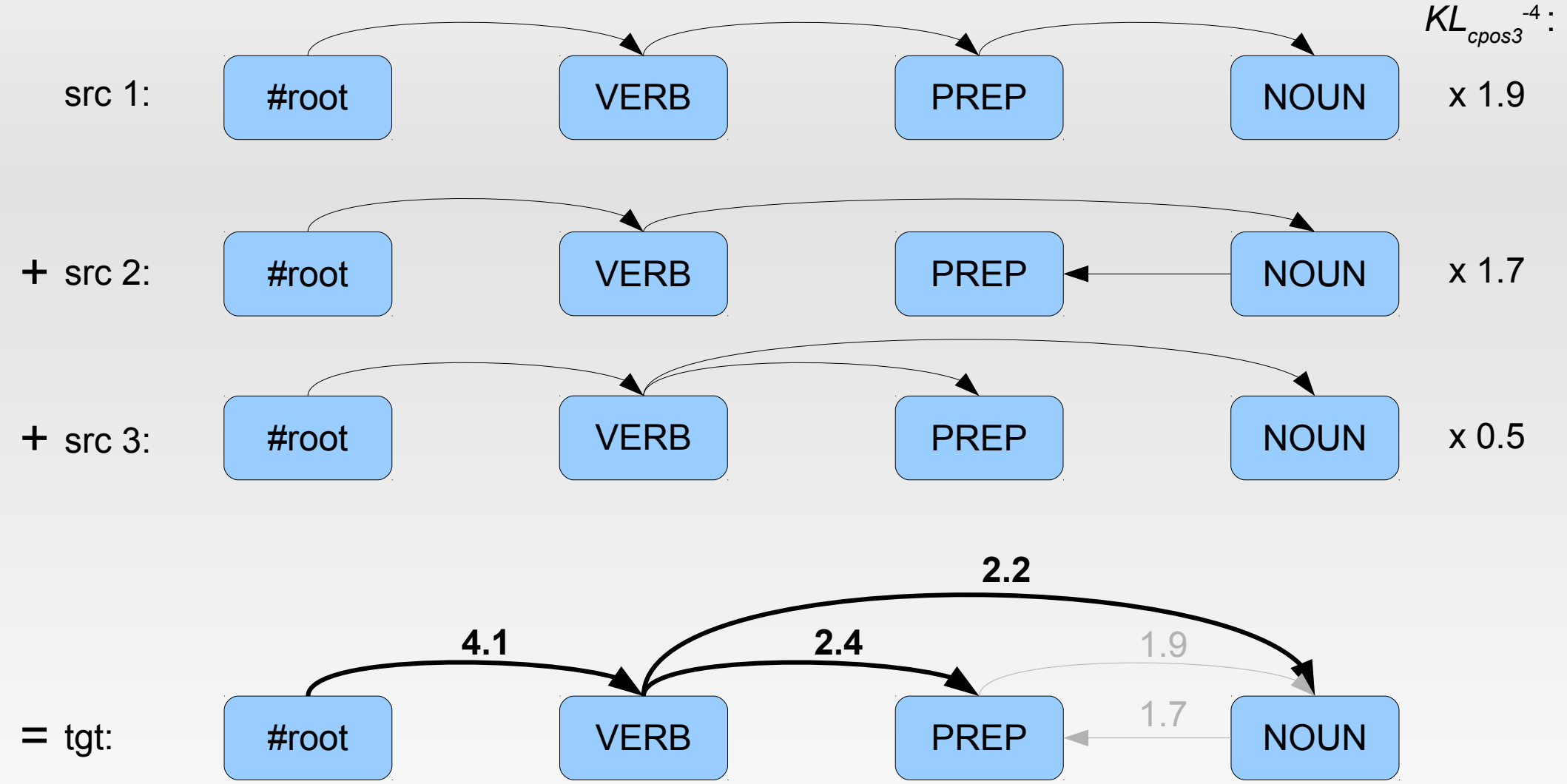
Parse tree combination



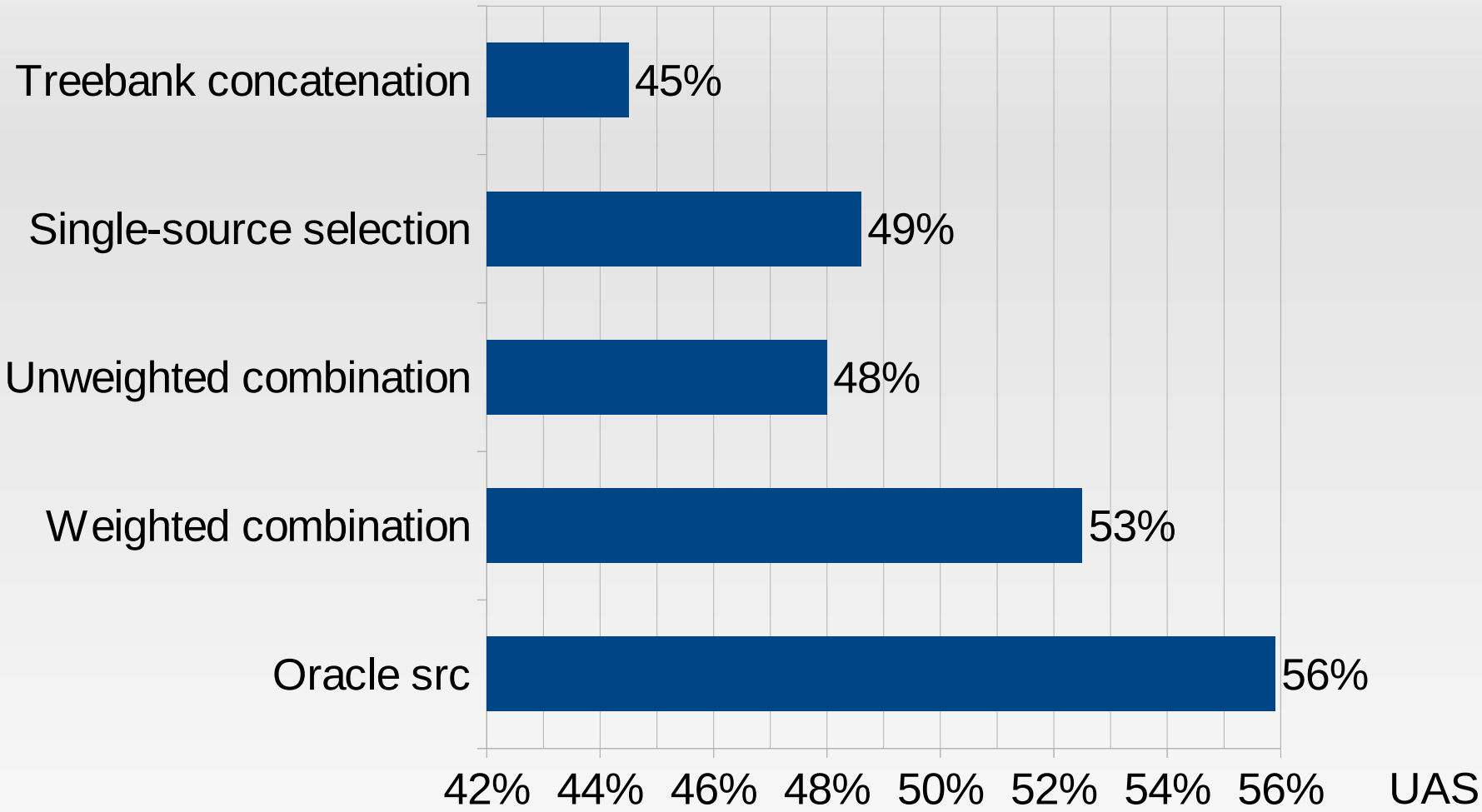
Weighted parse tree combination



Weighted parse tree combination



Average over 18 test TBs



Conclusion

- $KL_{cpos\ 3}$ language similarity measure
  - for delexicalized parser transfer
- single-source transfer
  - source treebank selection (44% success)
- multi-source transfer (tree combination)
  - source treebank weighting with $KL_{cpos\ 3}^{-4}$
  - +3.9% over single-source transfer
  - +4.5% over unweighted tree combination
  - +8.0% over treebank concatenation

Thank you for your attention

Rudolf Rosa, Zdeněk Žabokrtský
{rosa,zabokrtsky}@ufal.mff.cuni.cz

KL_{cpos^3}

a Language Similarity Measure for Delexicalized Parser Transfer

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

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