Rudolf Rosa
rosa@ufal.mff.cuni.cz

Defence of PhD thesis:

Discovering the structure of natural language sentences by semi-supervised methods

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

ÚFAL MFF UK, Praha, 14 June 2018

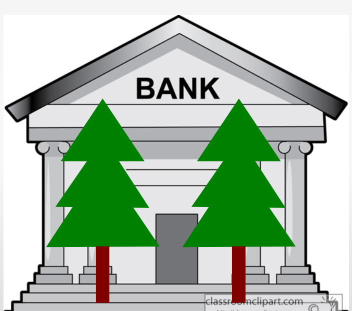
Syntactic analysis

Rudolf | likes | trains

Syntactic analysis

- take treebank (Universal Dependencies)

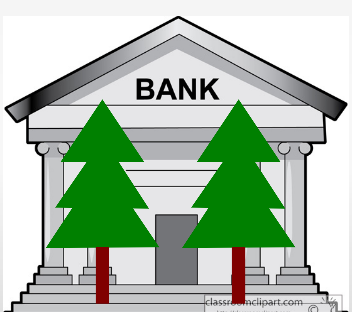
Rudolf likes trains

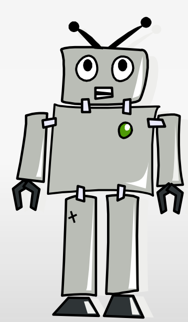


Syntactic analysis

- take treebank (Universal Dependencies)
- train tagger & parser (UDPipe)

Rudolf likes trains

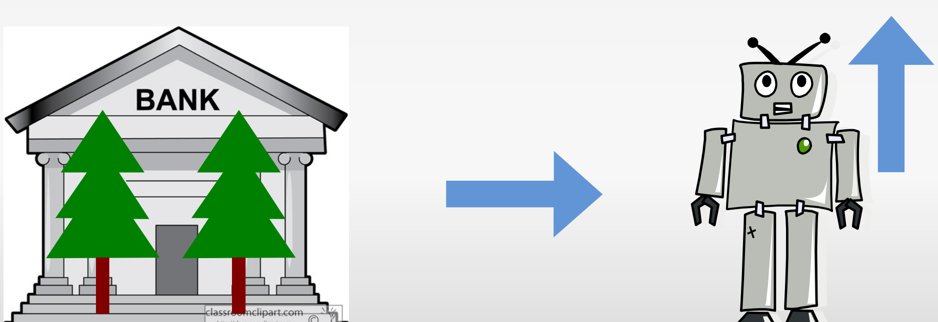




Syntactic analysis

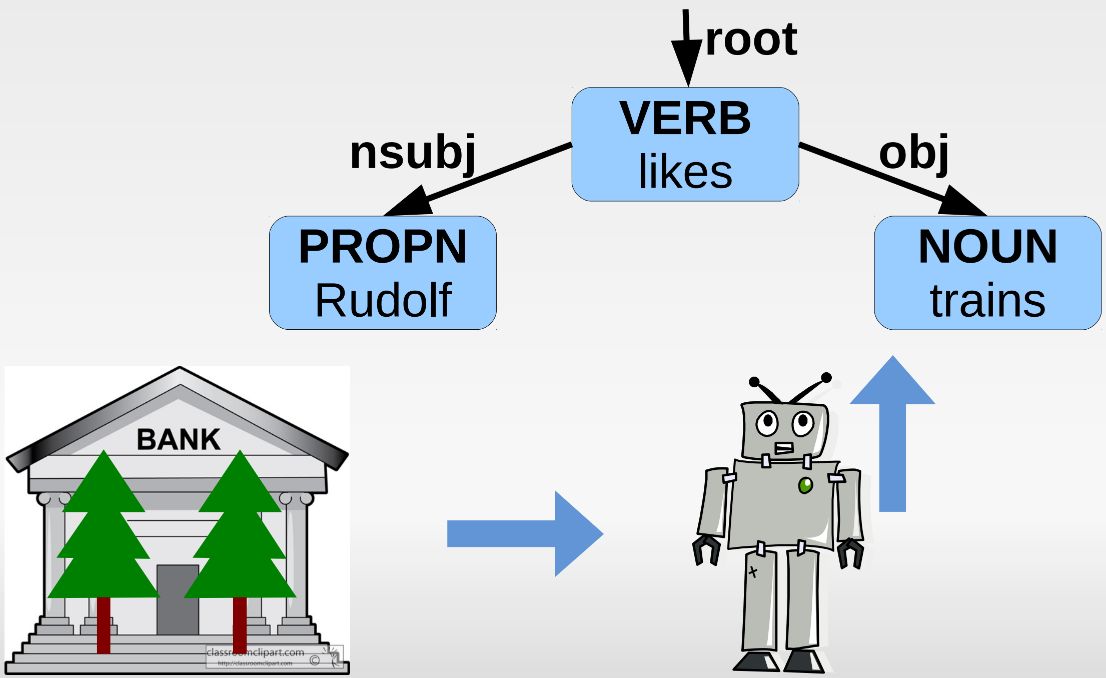
- take treebank (Universal Dependencies)
- train tagger & parser (UDPipe)
- tag (for universal part of speech)

PROPN	**VERB**	**NOUN**
Rudolf	likes	trains



Syntactic analysis

- take treebank (Universal Dependencies)
- train tagger & parser (UDPipe)
- tag (for universal part of speech)
- parse (into universal dependencies)



Under-resourced languages

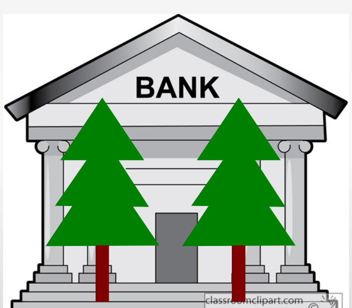
- take treebank...

Under-resourced languages

- take treebank...
  - ~70 languages covered (UD, HamleDT)

Under-resourced languages

- take treebank...
  - ~70 languages covered (UD, HamleDT)
  - ~7000 languages exist



Under-resourced languages

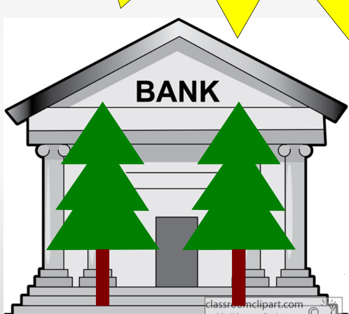
- take treebank...
  - ✓ ~70 languages covered (UD, HamleDT)
  - ✗ ~7000 languages exist

use
cross-lingual
approaches!

Under-resourced languages

- take treebank...
  - ~70 languages covered (UD, HamleDT)
  - ~7000 languages exist

use
cross-lingual
approaches!



Suggested approach

- parsing an under-resourced target language
  - i.e. we have no treebank for the target language
  - here simulated by Slovak

Rudolf ľúbi vlaky

Suggested approach

Rudolf

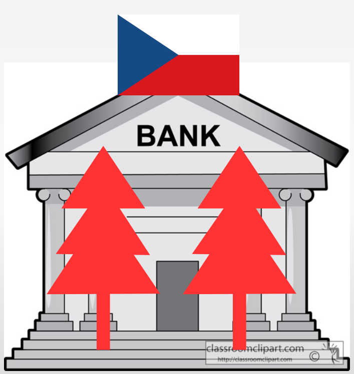
ľúbi

vlaky

Suggested approach

- take treebank for a source language
  - e.g. the Czech Prague Dependency Treebank

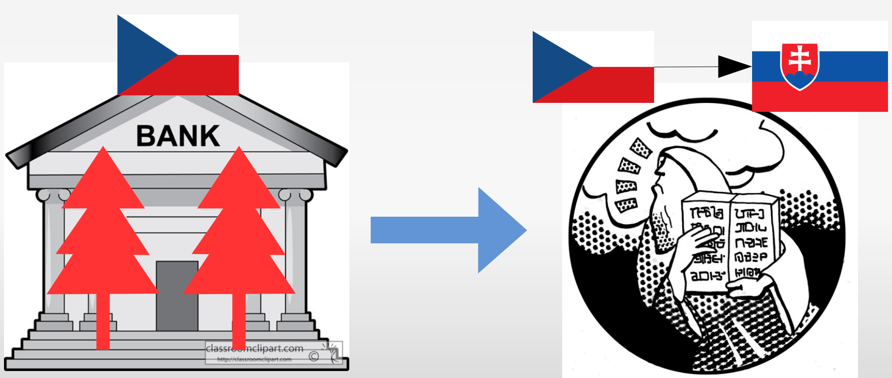
Rudolf ľúbi vlaky



Suggested approach

- take treebank for a <u>source</u> language
- machine translate into the <u>target</u> language

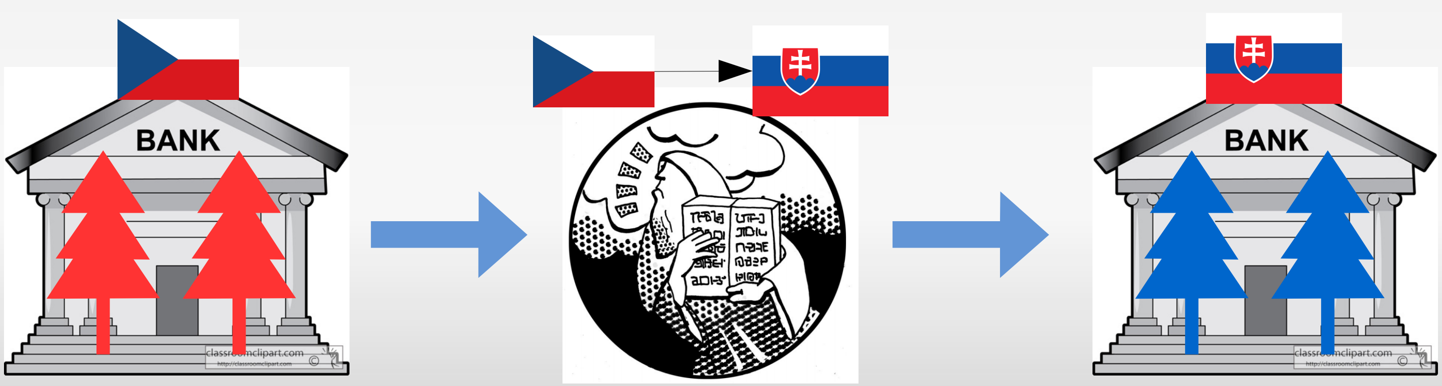
Rudolf ľúbi vlaky



Suggested approach

- take treebank for a source language
- machine translate into the target language
  - get a pseudo-target treebank

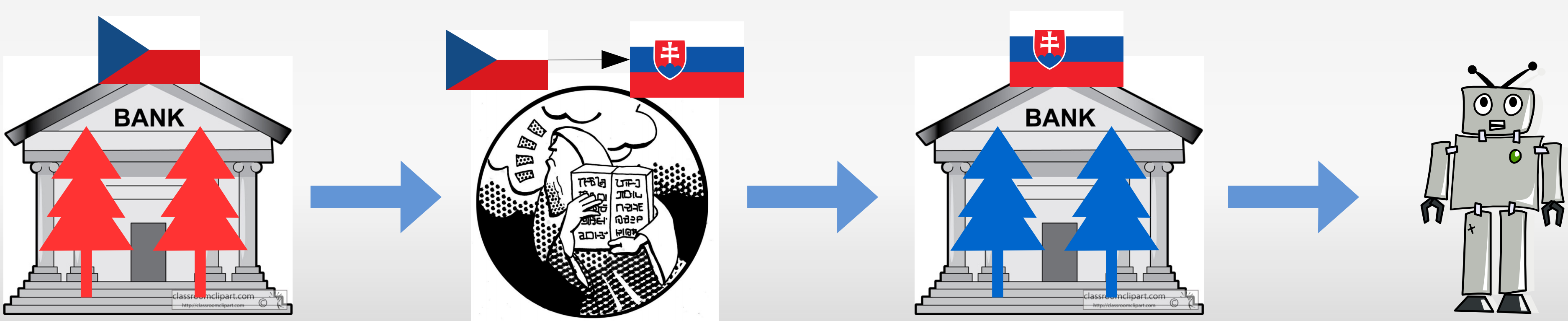
Rudolf ľúbi vlaky



Suggested approach

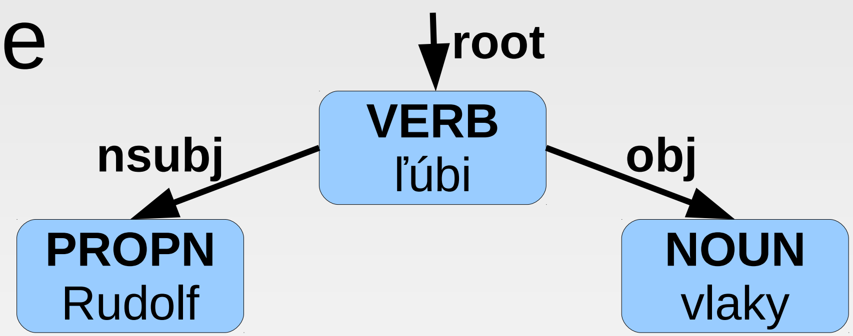
- take treebank for a source language
- machine translate into the target language
- train tagger & parser

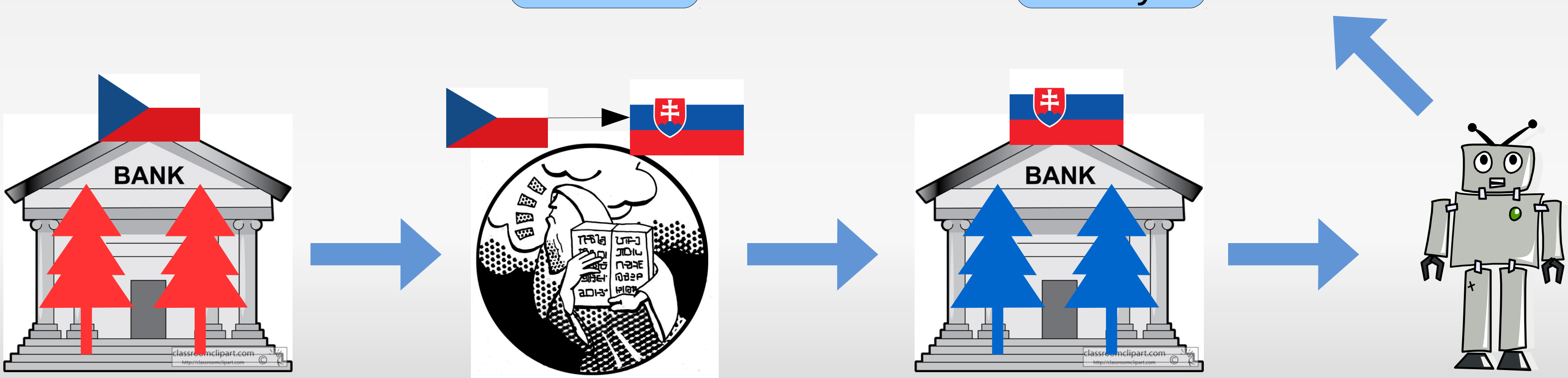
Rudolf ľúbi vlaky



Suggested approach

- take treebank for a source language
- machine translate into the target language
- train tagger & parser
- tag and parse



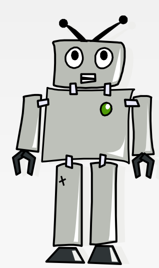


Problems to solve

- take treebank for a source language

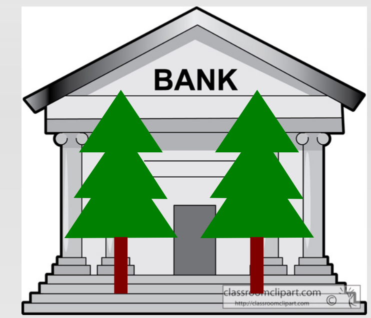
- machine translate into the target language

- train tagger & parser
- tag and parse



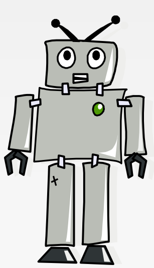
Problems to solve

- take treebank for a source language
  - how to choose the language?



- machine translate into the target language

- train tagger & parser
- tag and parse

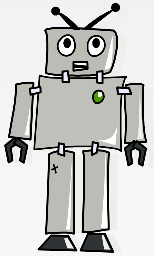


Problems to solve

- take treebank for a source language
  - how to choose the language?
  - combine multiple languages?

- machine translate into the target language

- train tagger & parser
- tag and parse

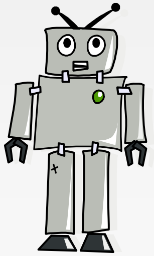


Problems to solve

- take treebank for a source language
  - how to choose the language?
  - combine multiple languages?

- machine translate into the target language
  - where to get the MT system?

- train tagger & parser
- tag and parse

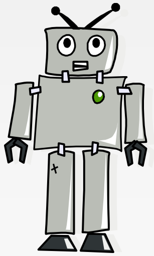


Problems to solve

- take treebank for a source language
  - how to choose the language?
  - combine multiple languages?

- machine translate into the target language
  - where to get the MT system?
  - how to translate a treebank?

- train tagger & parser
- tag and parse



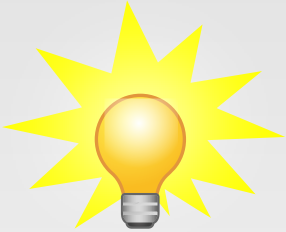
Choosing the source language

Choosing the source language

- source should be very similar to target language
  - family, word order, auxiliaries...

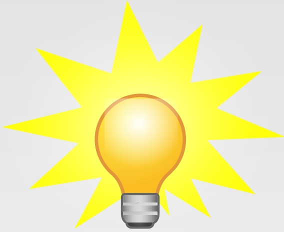
Choosing the source language

- source should be very similar to target language
  - family, word order, auxiliaries...
- use your linguistic knowledge&experience
  - does not scale well

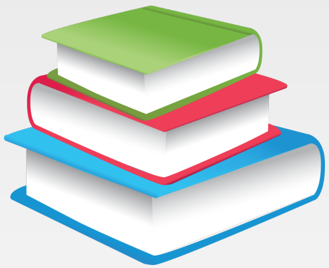


Choosing the source language

- source should be very similar to target language
  - family, word order, auxiliaries...
- use your linguistic knowledge&experience
  - does not scale well

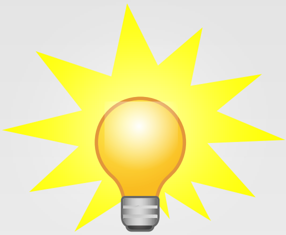


- look e.g. into the World Atlas of Language Structures
  - typological properties

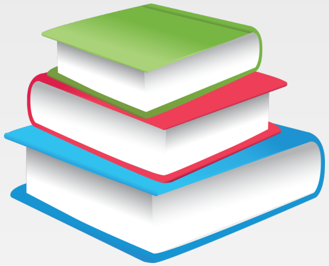


Choosing the source language

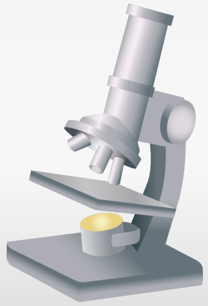
- source should be very similar to target language
  - family, word order, auxiliaries...
- use your linguistic knowledge&experience
  - does not scale well



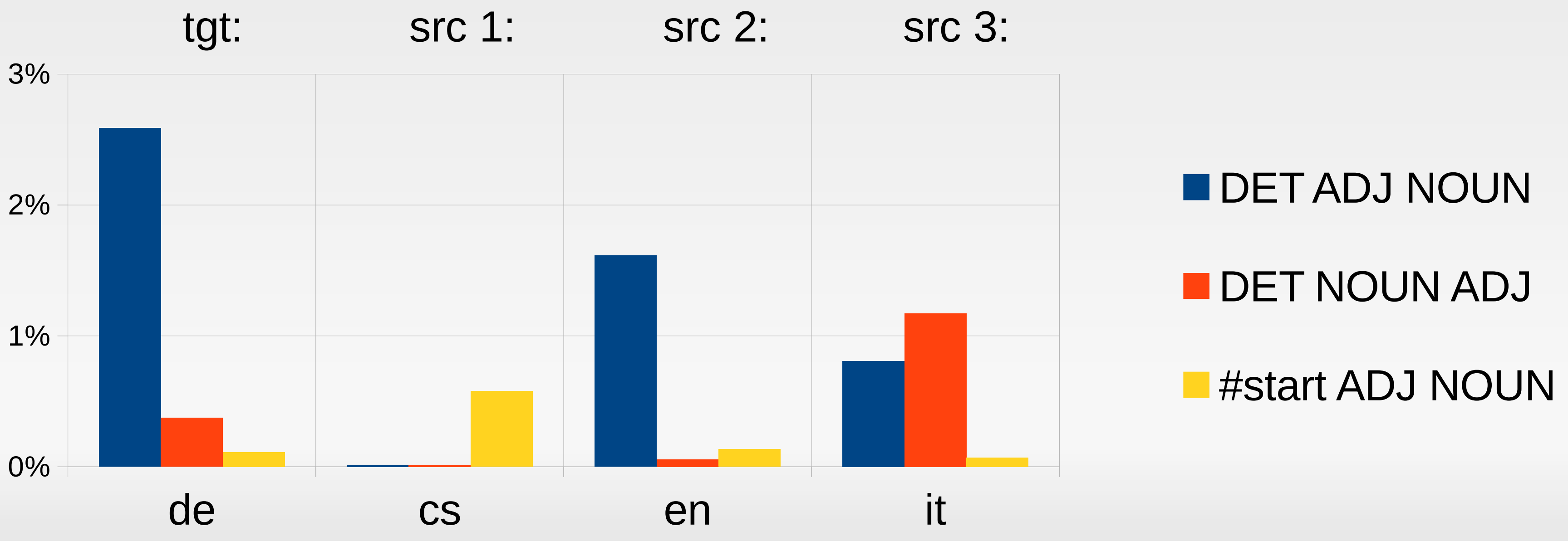
- look e.g. into the World Atlas of Language Structures
  - typological properties



- look at part-of-speech tags
  - empirical measures



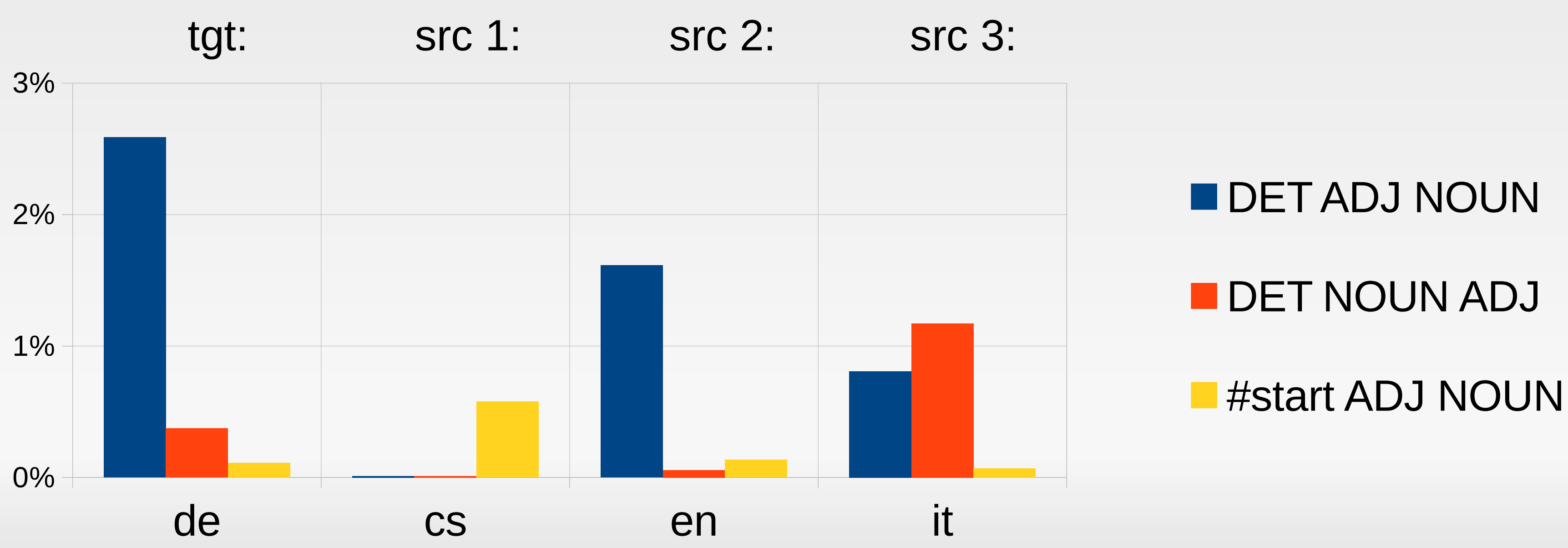
KL_{cpos^3} language similarity



KL_{cpos^3} language similarity

- Kullback-Leibler divergence of POS trigram distributions

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



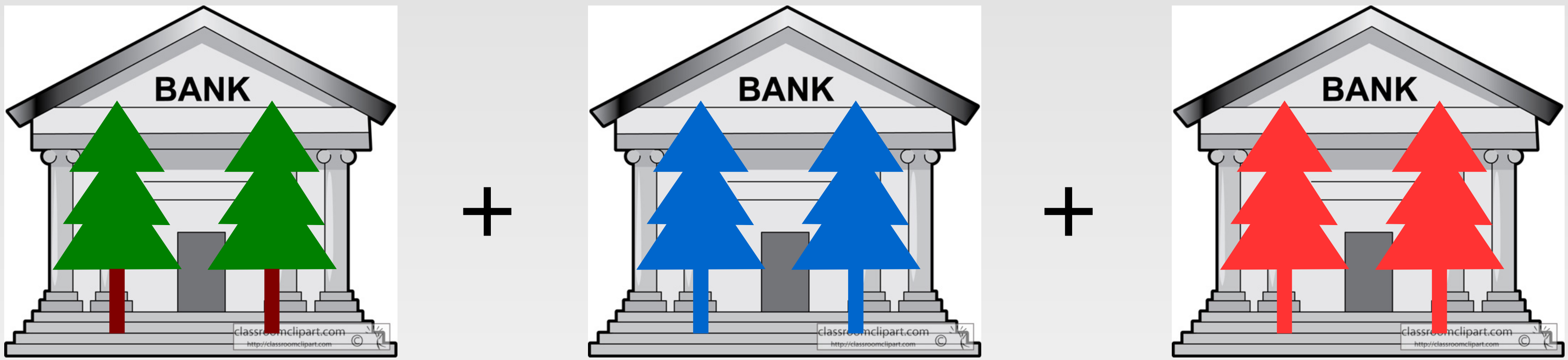
KL_{cpos^3} language similarity

- good performance
  - independently confirmed (Agić, 2017)
  - identifies best source treebank in ~50% cases

KL_{cpos^3} language similarity

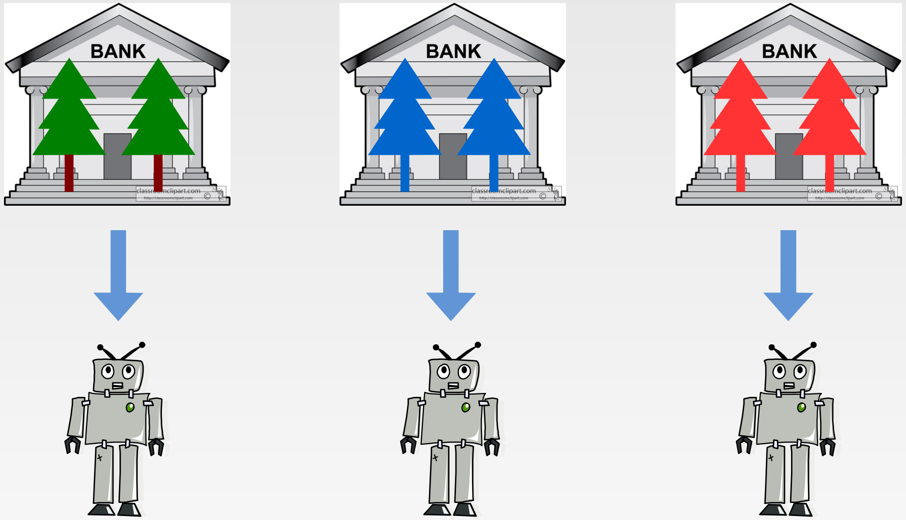
- good performance
  - independently confirmed (Agić, 2017)
  - identifies best source treebank in ~50% cases
- requires POS-tagged target data
  - developed with gold POS
  - also good with cross-lingual POS

Combining multiple sources



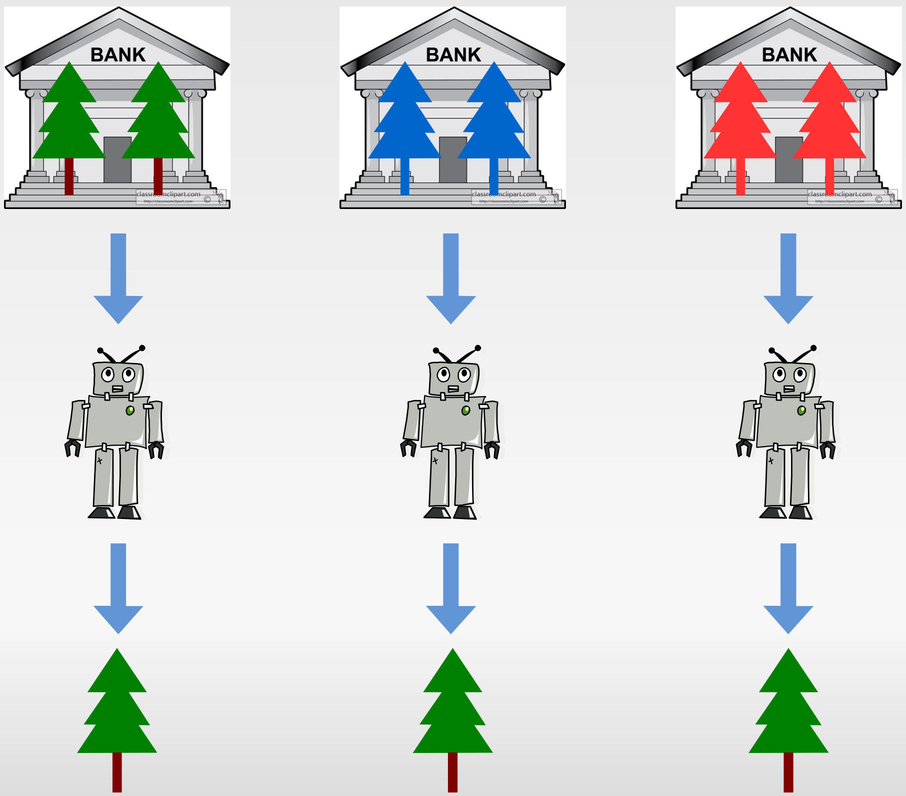
Parse tree combination

- train tagger&parser on each translated treebank



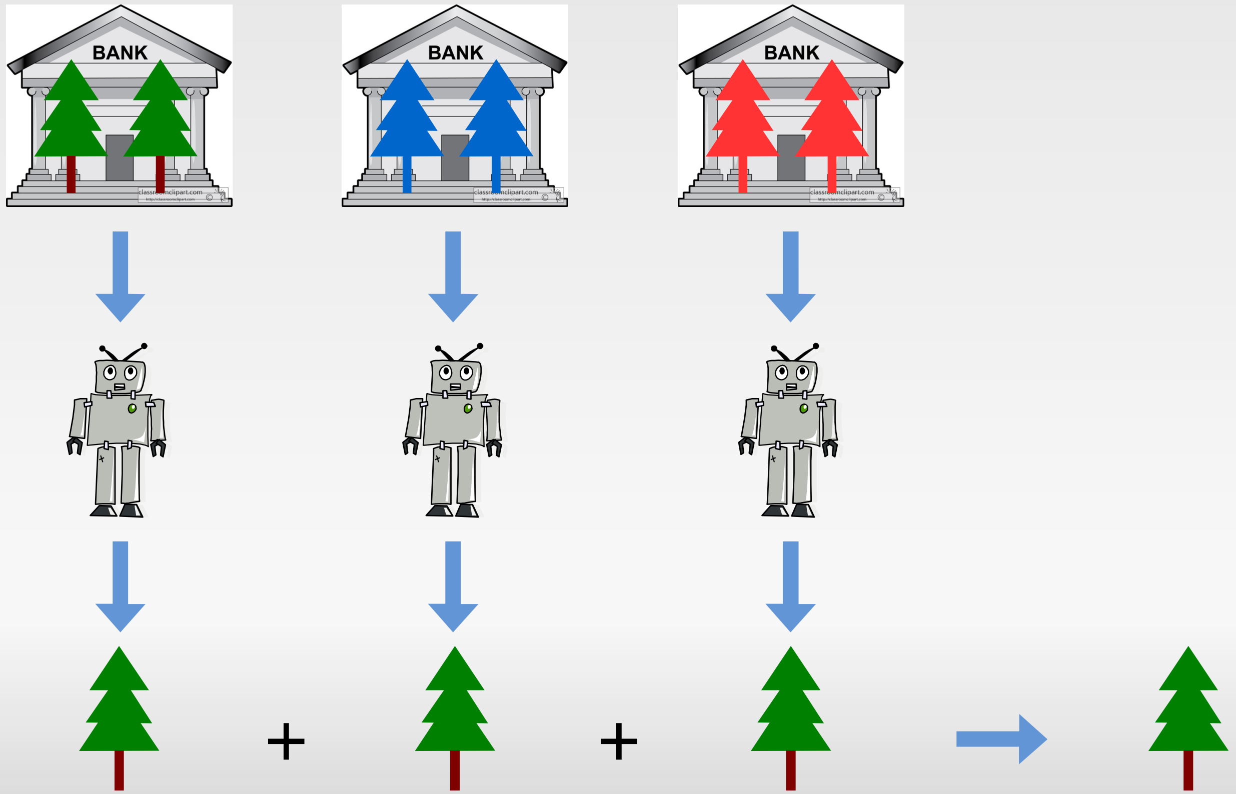
Parse tree combination

- train tagger&parser on each translated treebank
- separately apply to target text



Parse tree combination

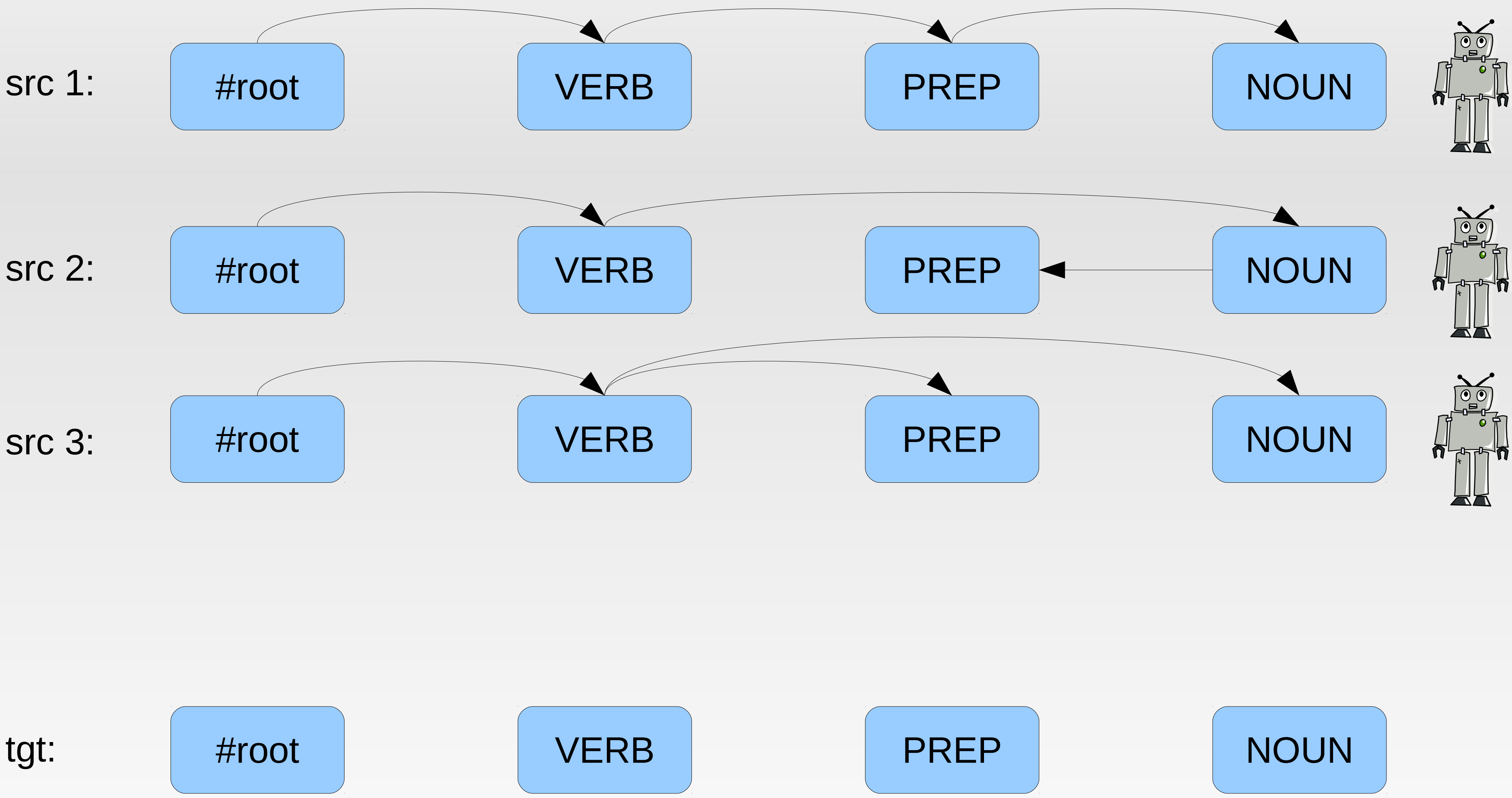
- train tagger&parser on each translated treebank
- separately apply to target text
- combine the resulting parse trees with MST



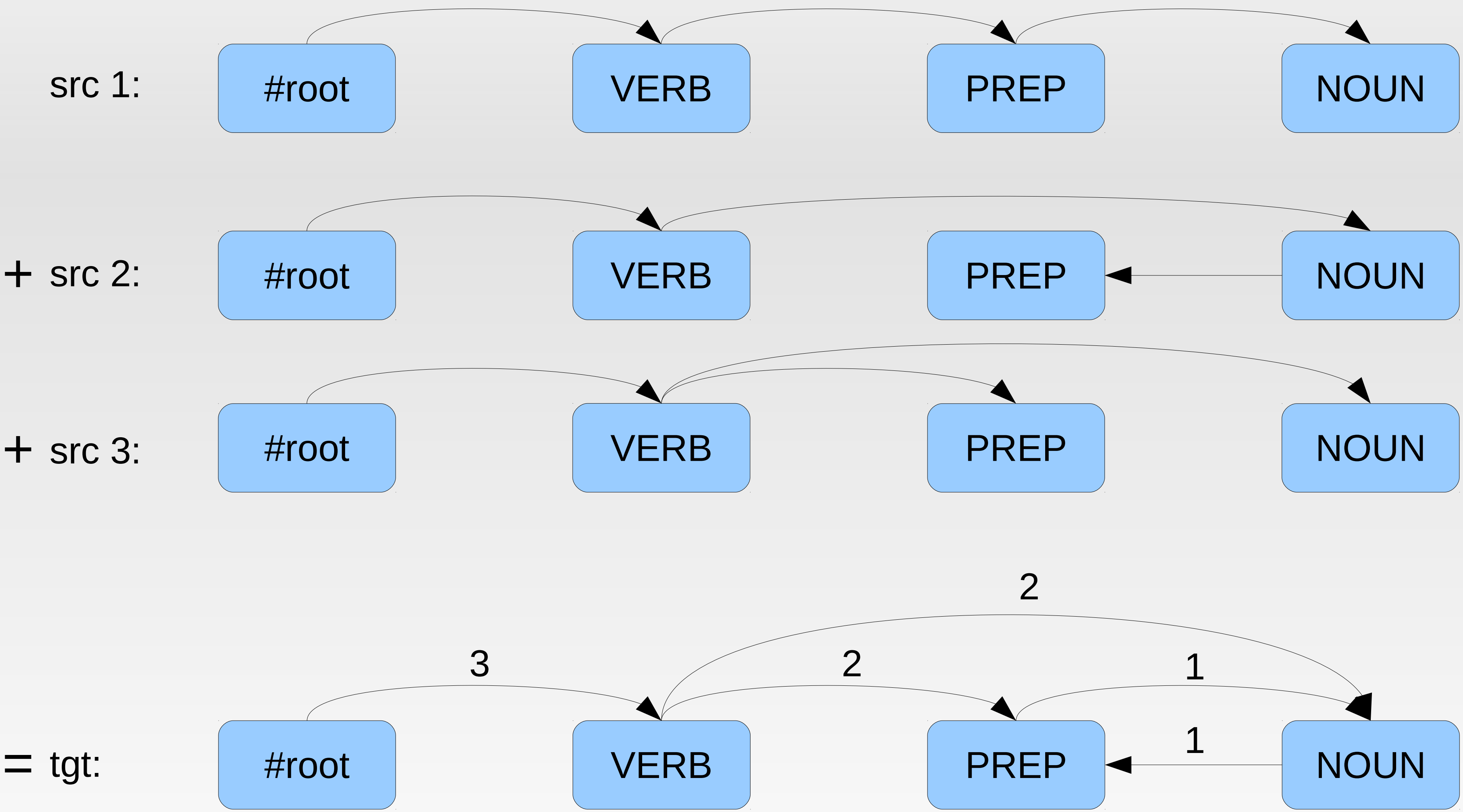
Parse tree combination

tgt: #root VERB PREP NOUN

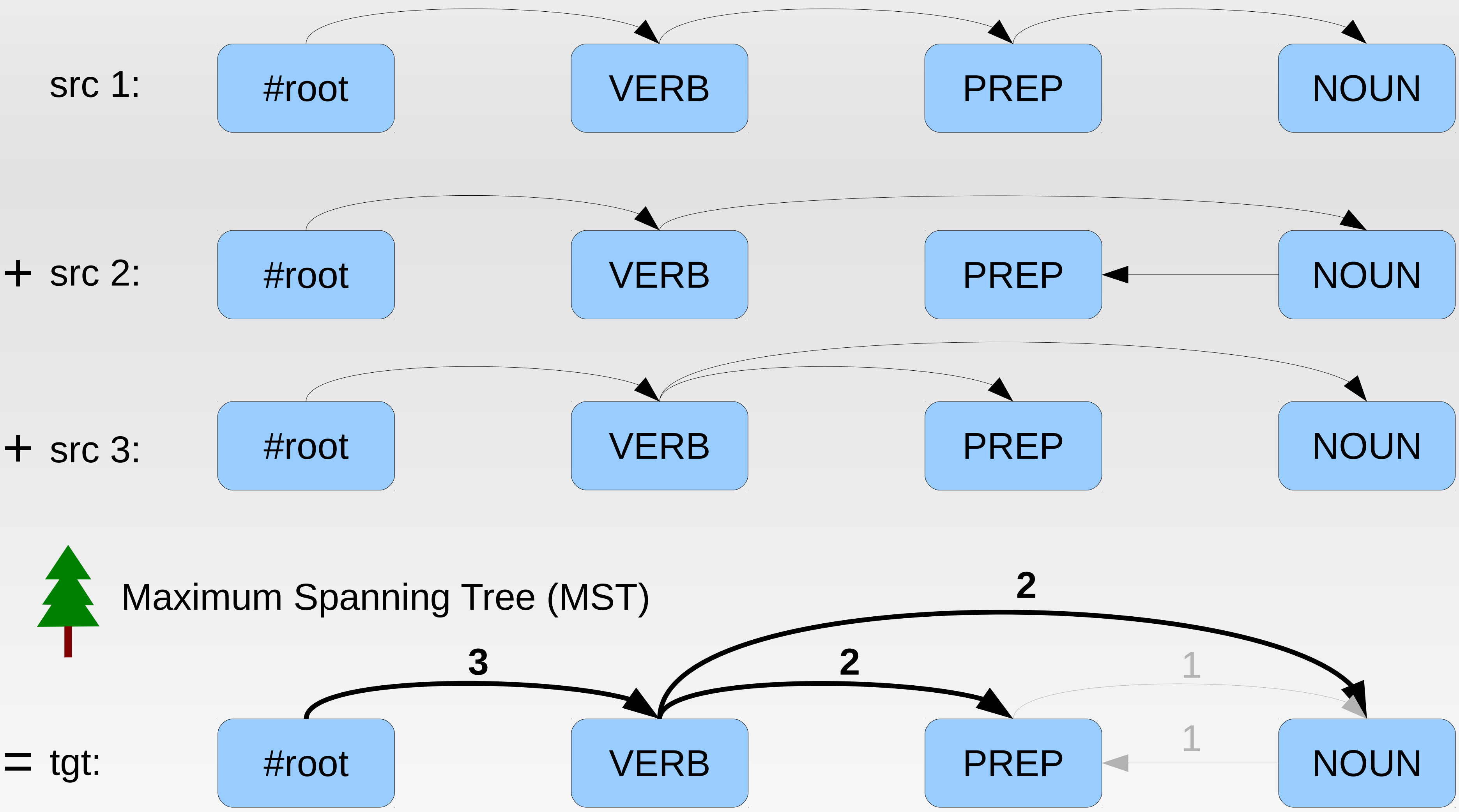
Parse tree combination



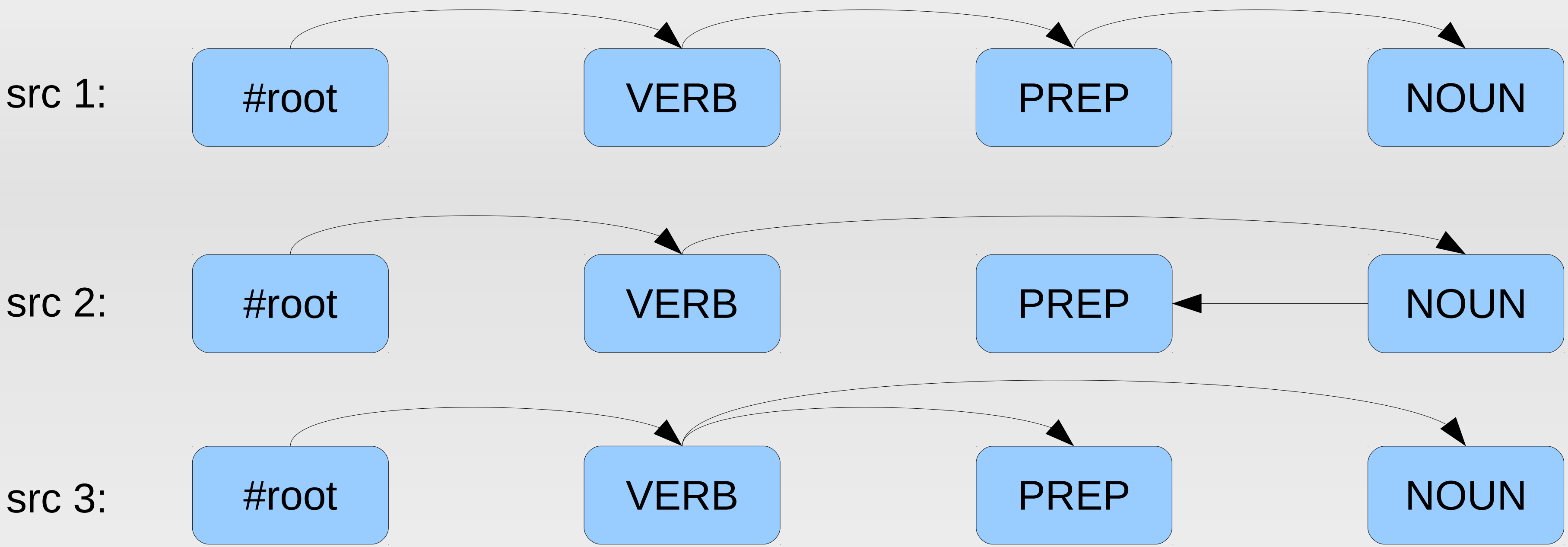
Parse tree combination

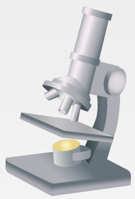


Parse tree combination

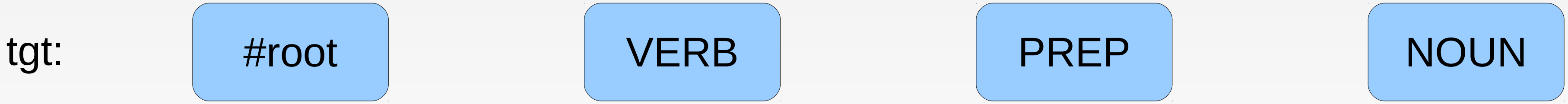


Weighted parse tree combination

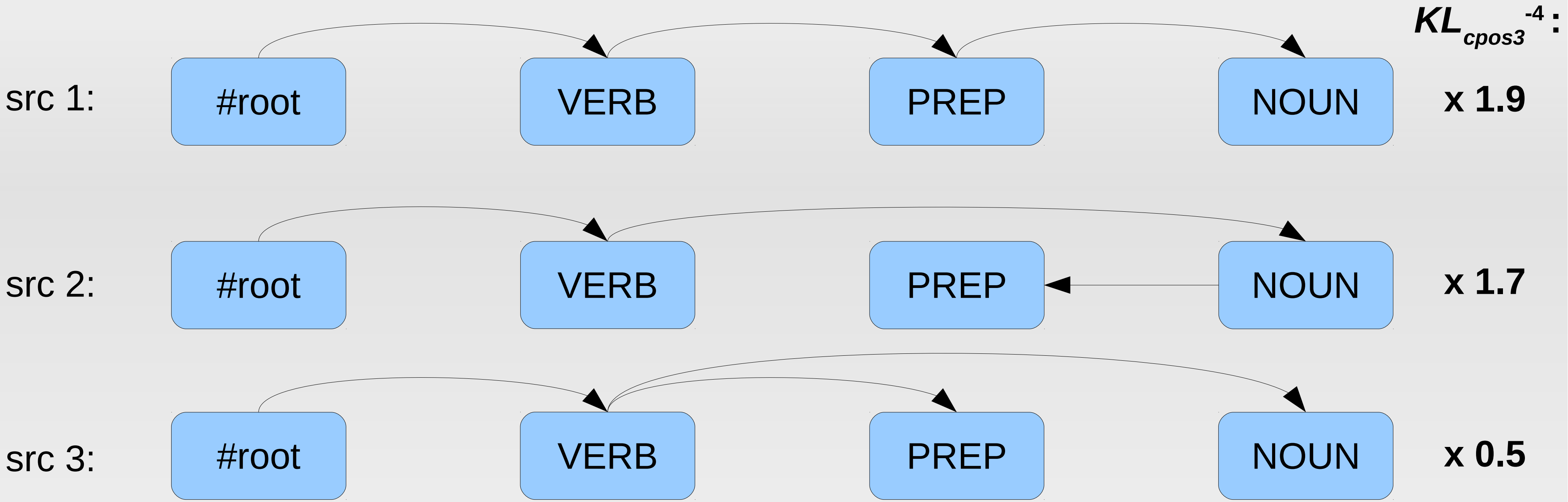


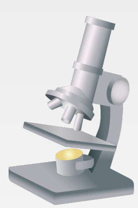


KL_{cpos3} distance → KL_{cpos3}^{-4} similarity

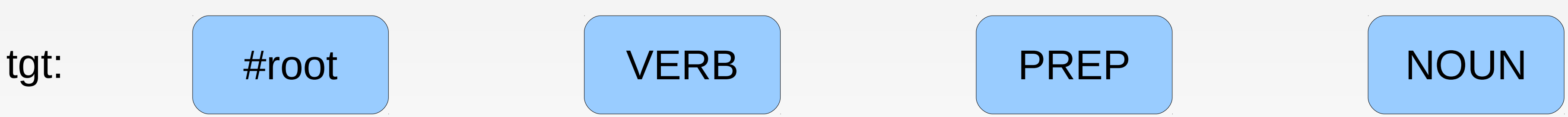


Weighted parse tree combination

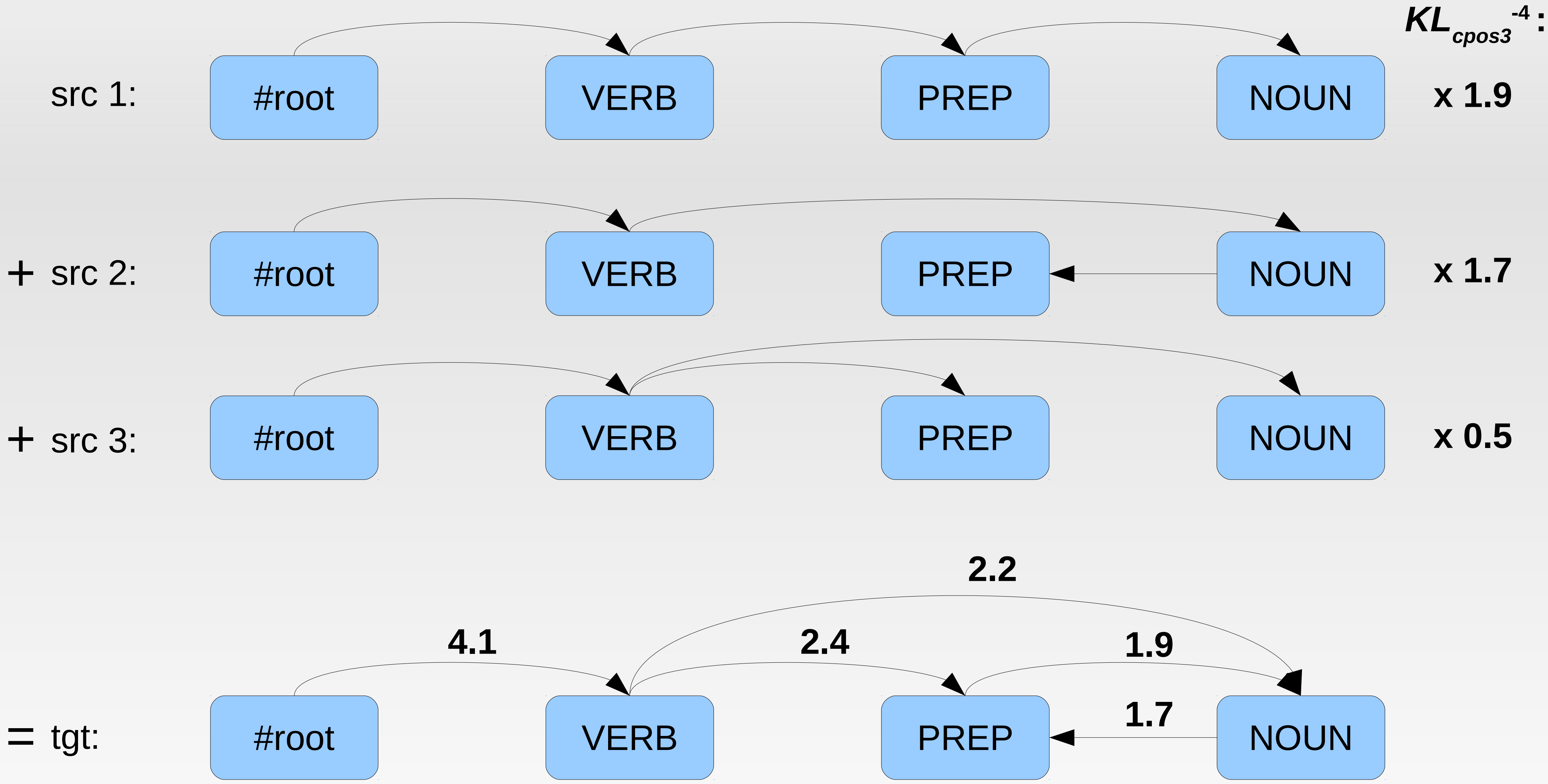


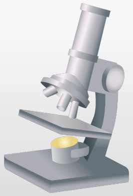


KL_{cpos3} distance → KL_{cpos3}^{-4} similarity



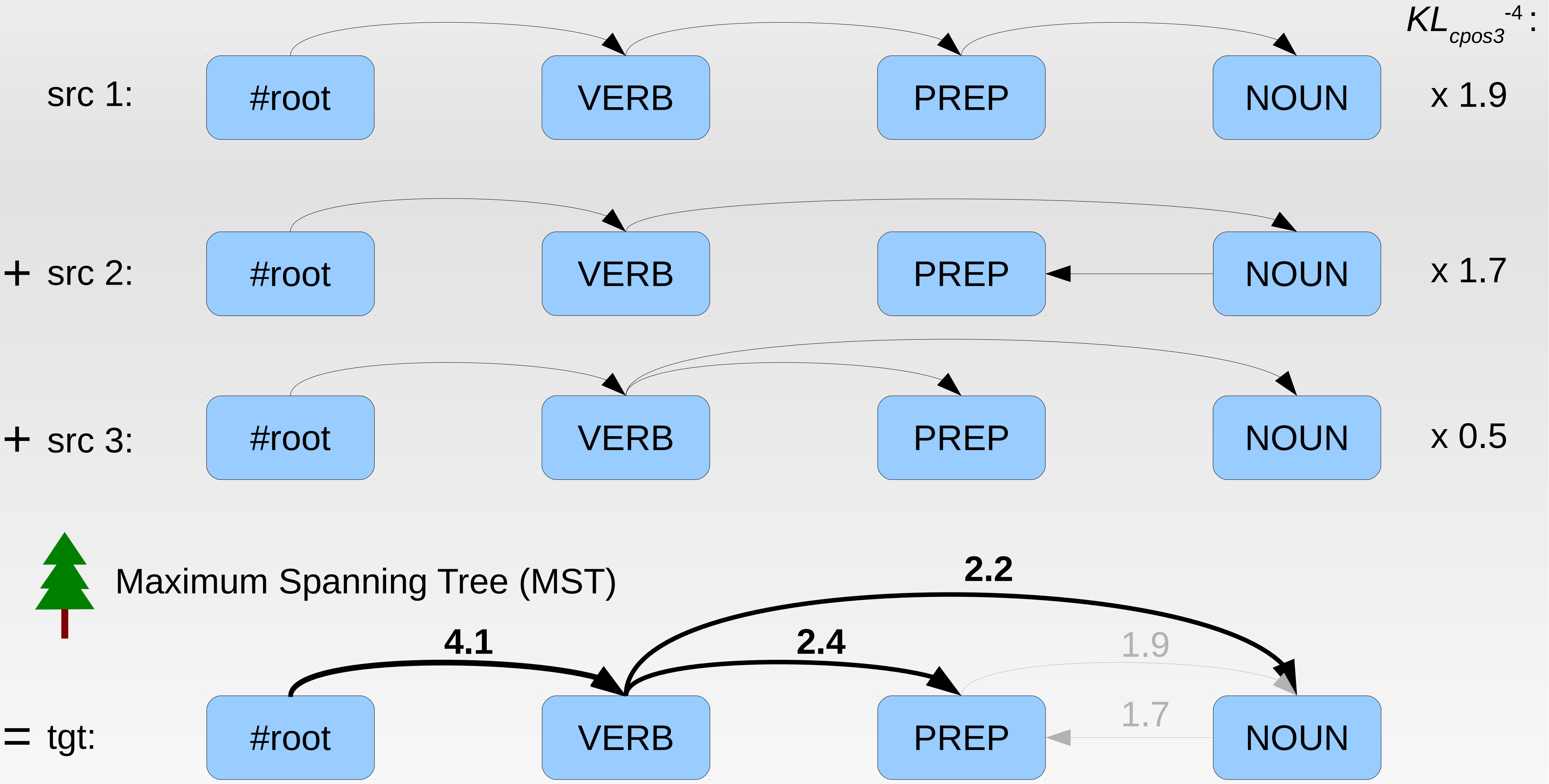
Weighted parse tree combination





KL_{cpos3} distance → KL_{cpos3}^{-4} similarity

Weighted parse tree combination



KL_{cpos3} distance → KL_{cpos3}^{-4} similarity

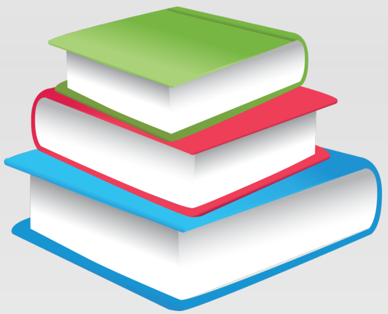
Translating the treebanks

Where to get the MT system

- train on source-target parallel texts

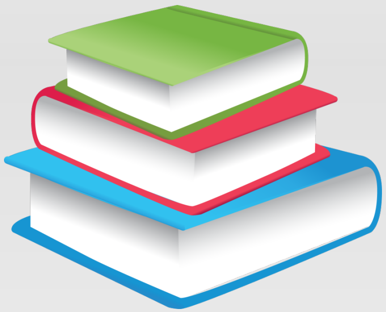
Where to get the MT system

- train on source-target parallel texts
- resource-rich languages: lots of data
  - laws, websites, literature, film subtitles...



Where to get the MT system

- train on source-target parallel texts
- resource-rich languages: lots of data
  - laws, websites, literature, film subtitles...

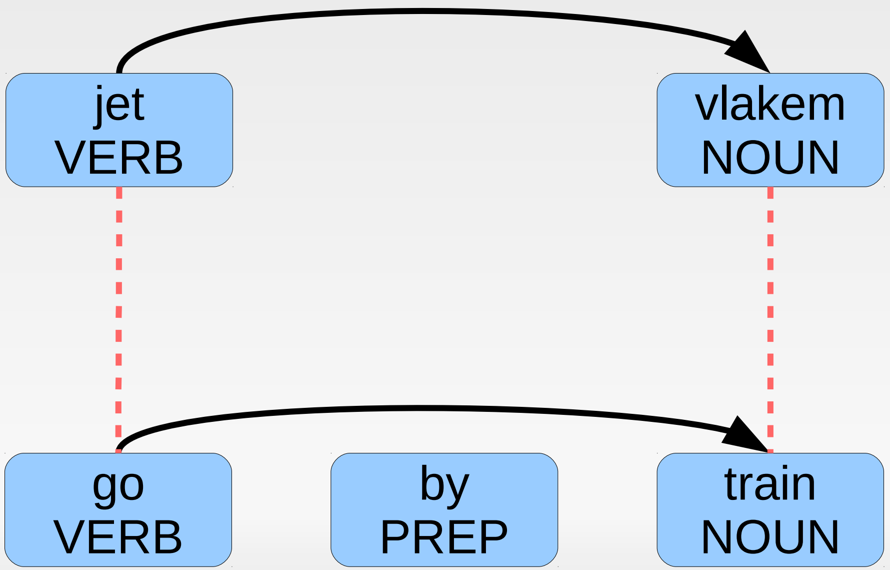


- resource-poor languages

  - often also poor on parallel data
  - Bible: 10k sentences for ~1200 languages
  - Watchtower: 100k sentences for ~135 languages

How to translate

- source and target sentences do not map 1:1
  - problems even with very similar languages
  - obviously worse for more distant languages



Solutions to non-isomorphism

Solutions to non-isomorphism

- ignore it, use 1:1 alignment
  - Moses with phrase length = 1 (± reordering)
  - lower-quality MT (in BLEU), but more literal → good

Tiedemann+ (2014), Rosa+ (2017)

Solutions to non-isomorphism

- ignore it, use 1:1 alignment

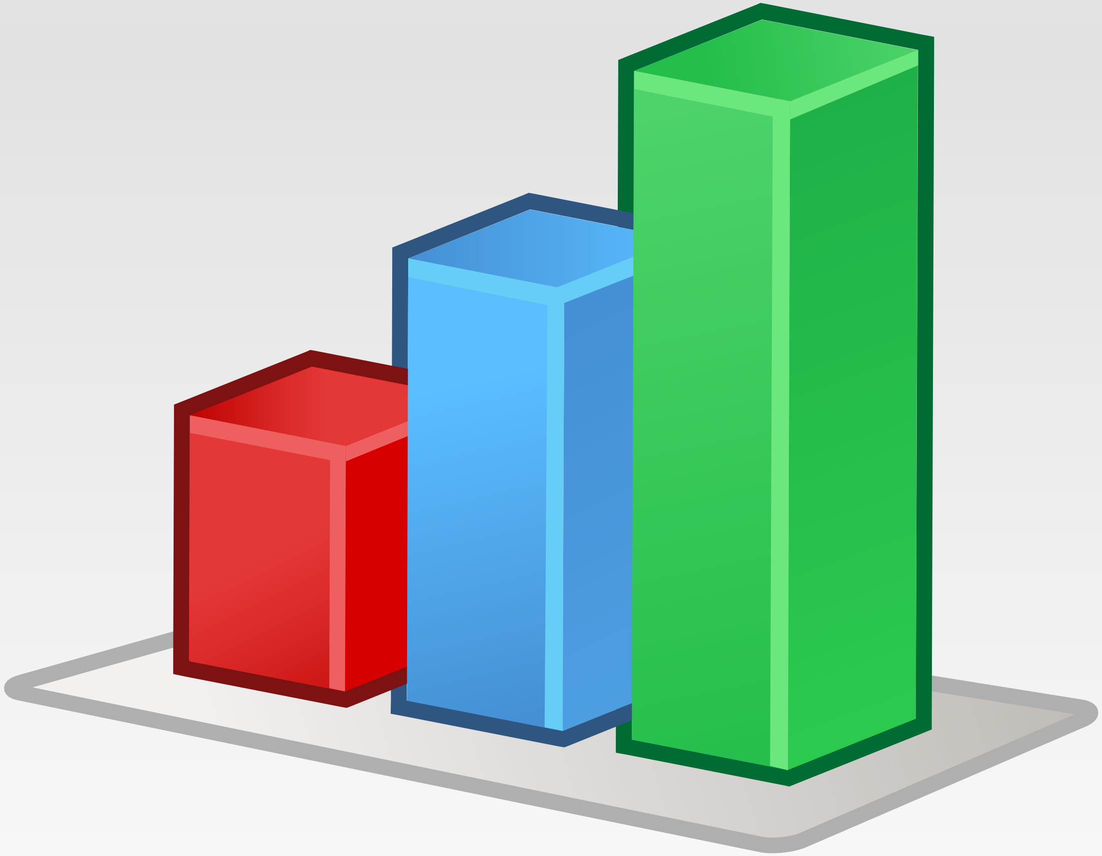
  Tiedemann+ (2014), Rosa+ (2017)

  - Moses with phrase length = 1 (± reordering)
  - lower-quality MT (in BLEU), but more literal → good
- complex projection heuristics

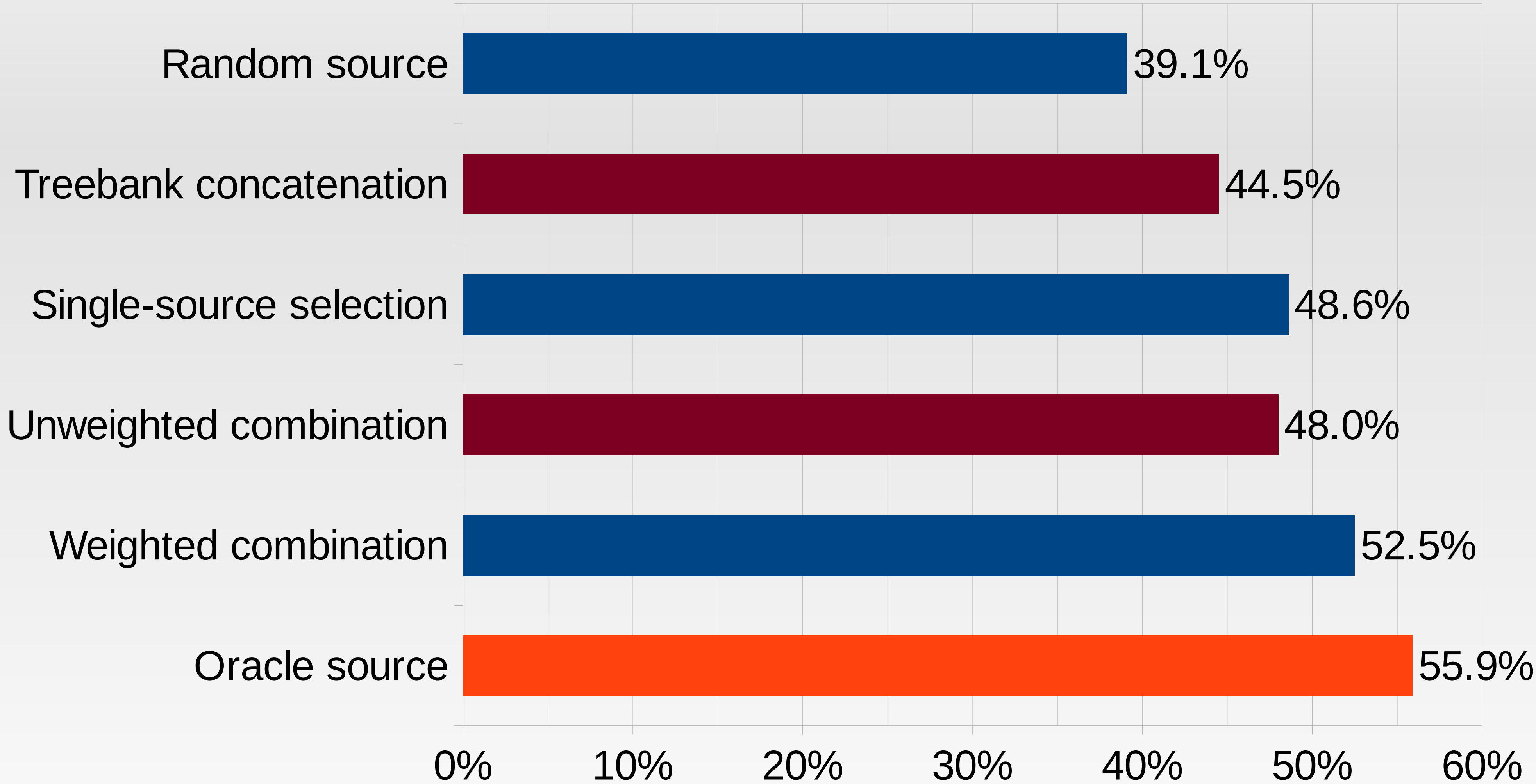
  Hwa+ (2005), Ramasamy (2014), Tiedemann+ (2014)

  - use M:N word-alignment and e.g. phrase-based MT
  - omit some nodes, guess some edges&deprels...
  - higher-quality MT (in BLEU), but projection noisy
  - for close languages performs similarly to word-based
  - for distant languages probably better

Evaluation

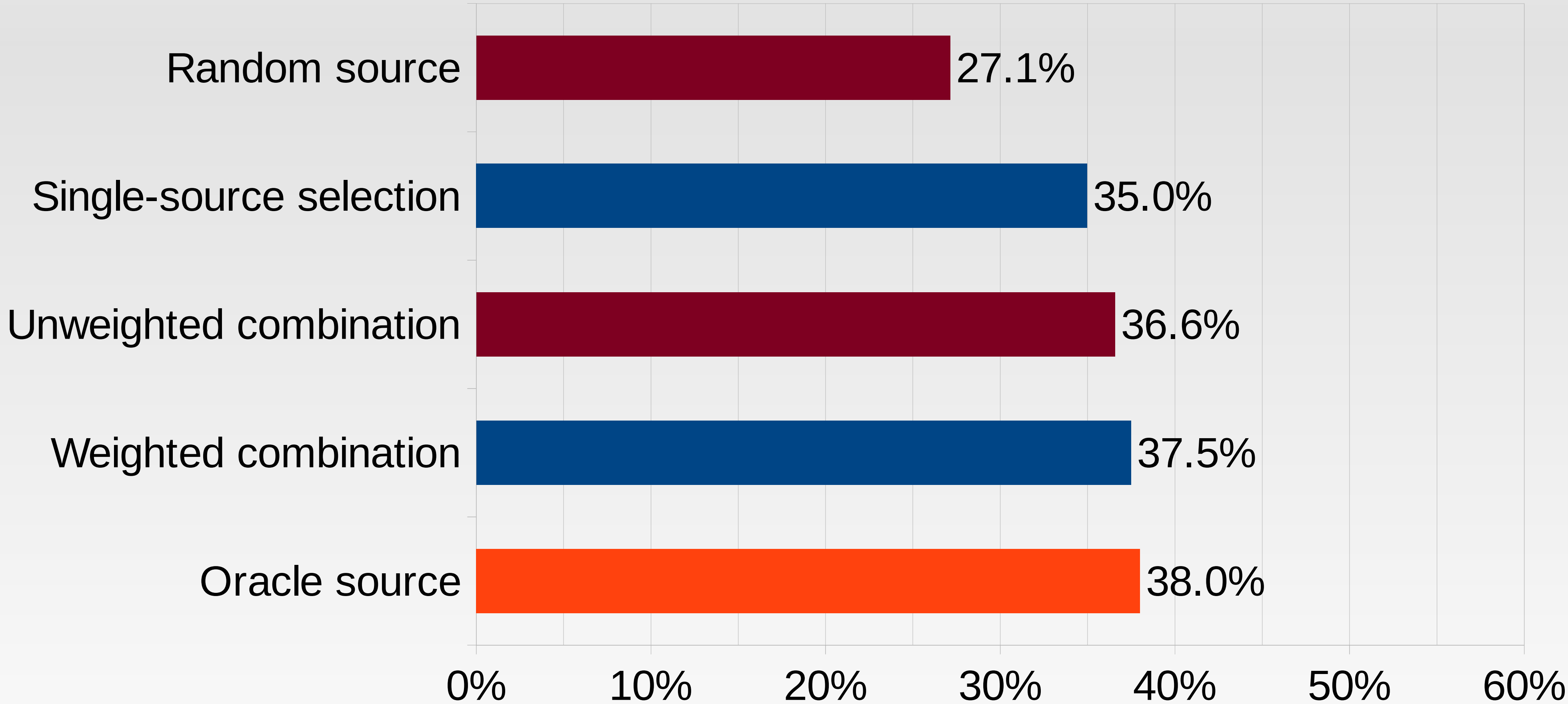


Source selection/combination



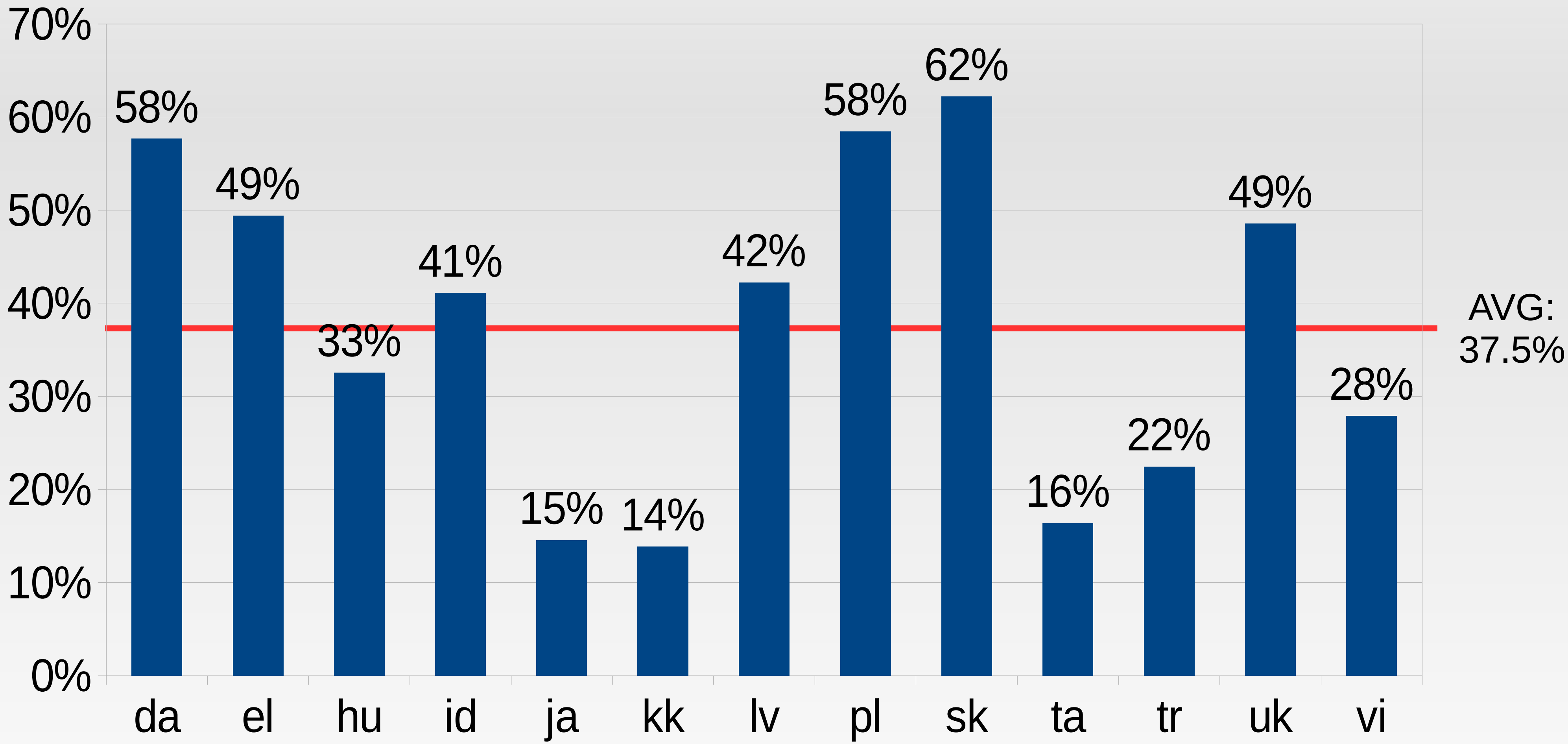
Unlabelled accuracy (UAS), gold POS, average over 18 target languages (HamleDT)

Source selection/combination



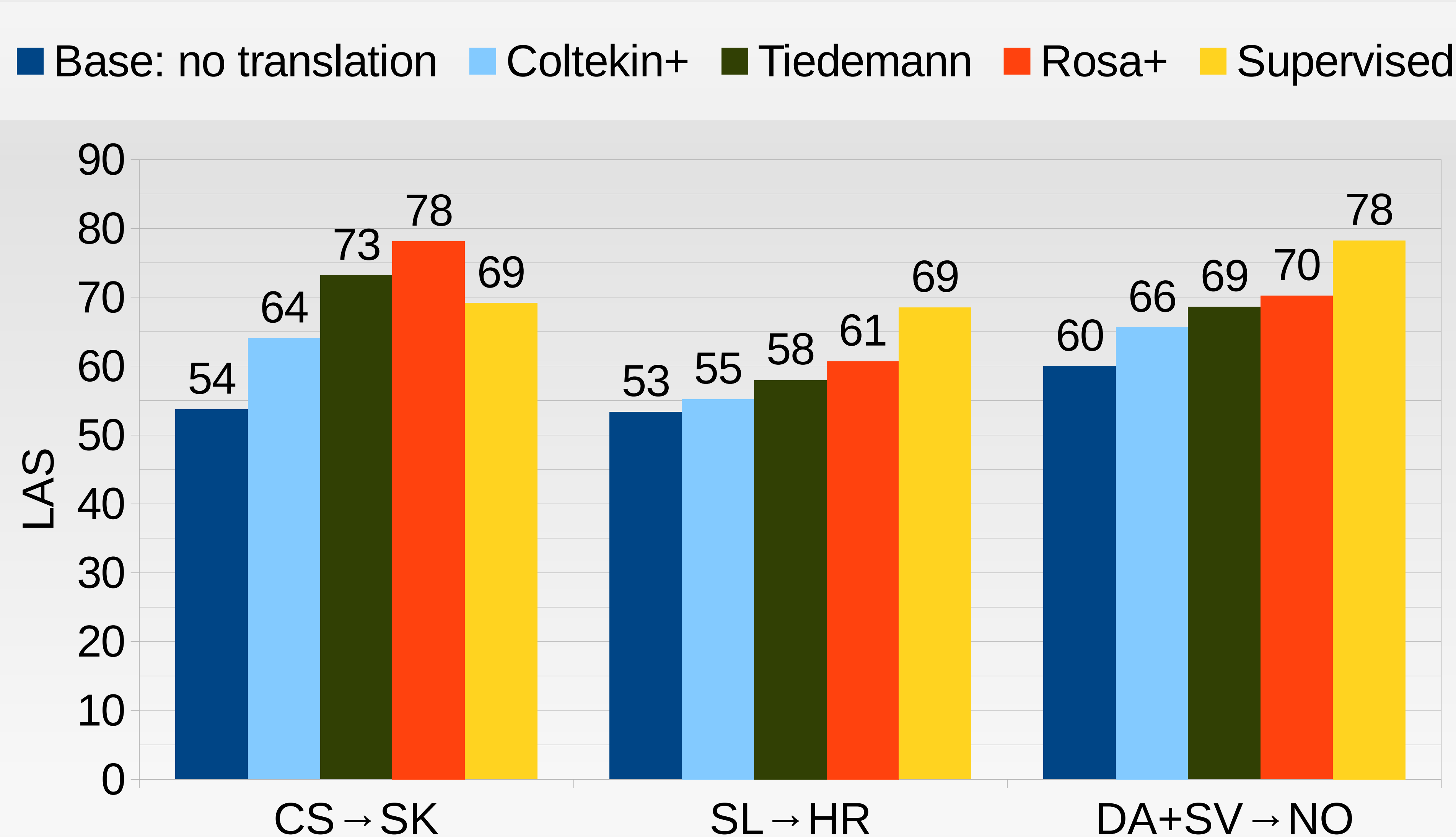
Labelled accuracy (LAS), cross-lingual POS, average over 13 target languages (UD)

Weighted combination



Labelled accuracy (LAS), cross-lingual POS, UD languages, weighted combination

VarDial 2017 shared task



Labelled accuracy (LAS), supervised POS, single source/concatenation

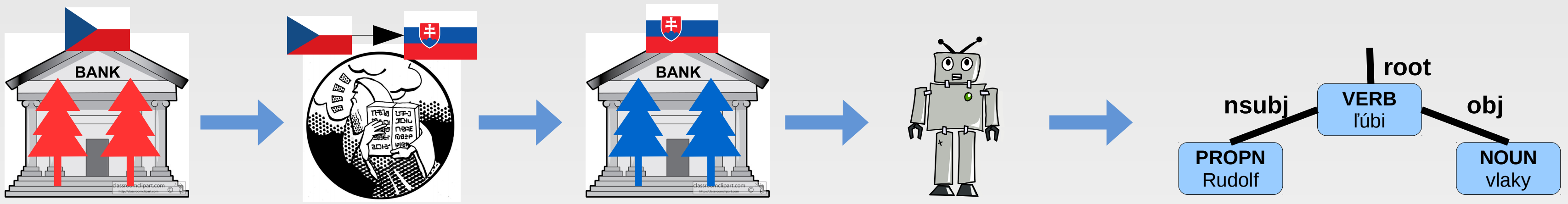
Conclusion

Conclusion

- Tagging and parsing of low-resourced languages

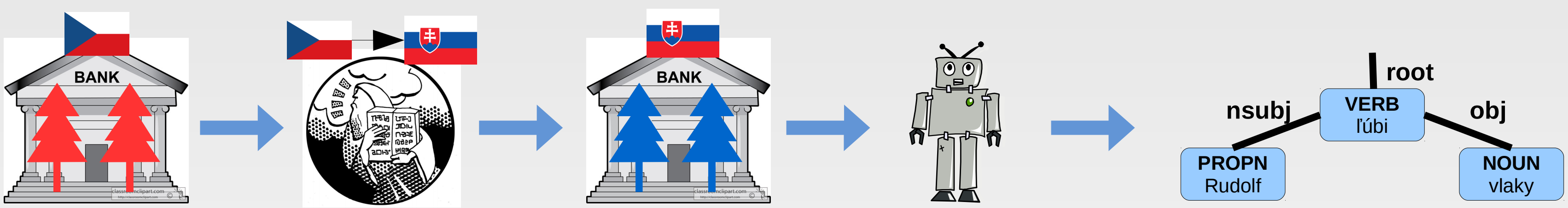
Conclusion

- Tagging and parsing of low-resourced languages
- Machine translation of source treebank
  - word-based monotone statistical machine translation

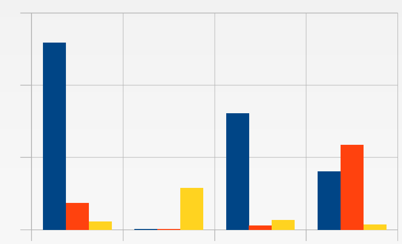


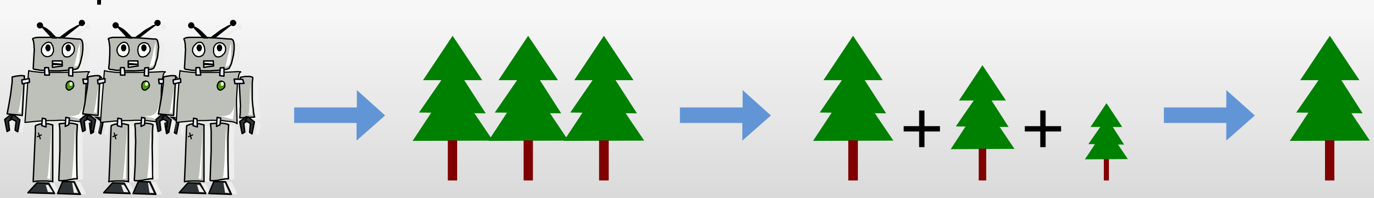
Conclusion

- Tagging and parsing of low-resourced languages
- Machine translation of source treebank
  - word-based monotone statistical machine translation



- Multiple sources available → select or combine
  - KL_{cpos3} language similarity (UPOS trigrams)
  - KL_{cpos3}^{-4} weighted parse tree combination





Thank you for your attention

Rudolf Rosa

rosa@ufal.mff.cuni.cz

Discovering the structure of natural language sentences by semi-supervised methods

Charles University

Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics

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

Q&A

1. NMT or similar approaches which can be inherently multilingual
2. looking at treebank size
3. left-arc and right-arc scores as edge weights
4. employing larger monolingual texts (in MT)

1. multilingual NMT/embeddings

- multilingual word embeddings
  - preliminary experiments (in 2016): worse than MT

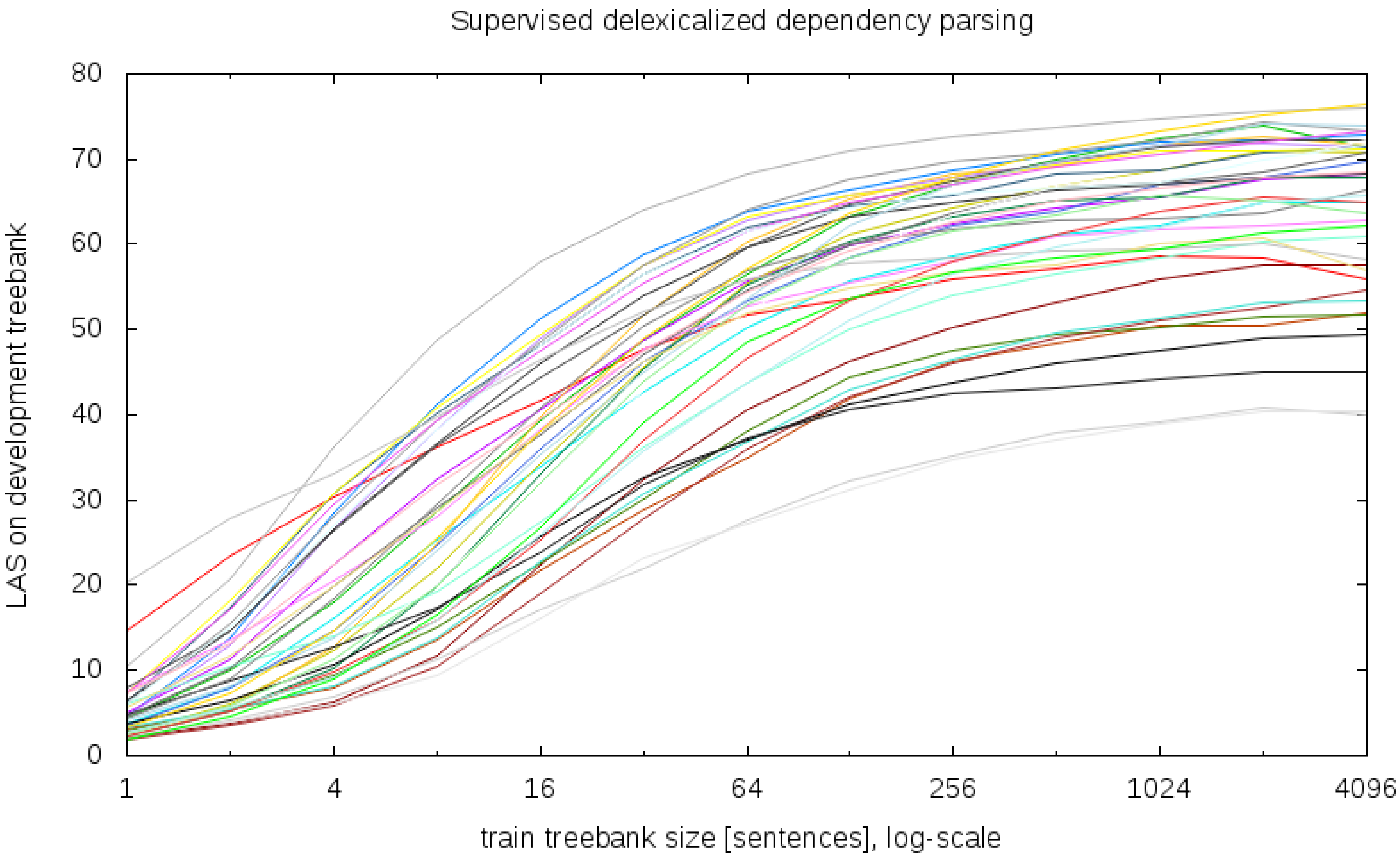
1. multilingual NMT/embeddings

- multilingual word embeddings
  - preliminary experiments (in 2016): worse than MT
- multilingual neural machine translation
  - have not tried, but looks promising, keen to try
  - zero-shot: jointly train A→B & B→C, translate A→C
    - results not terrific, but probably worth trying out

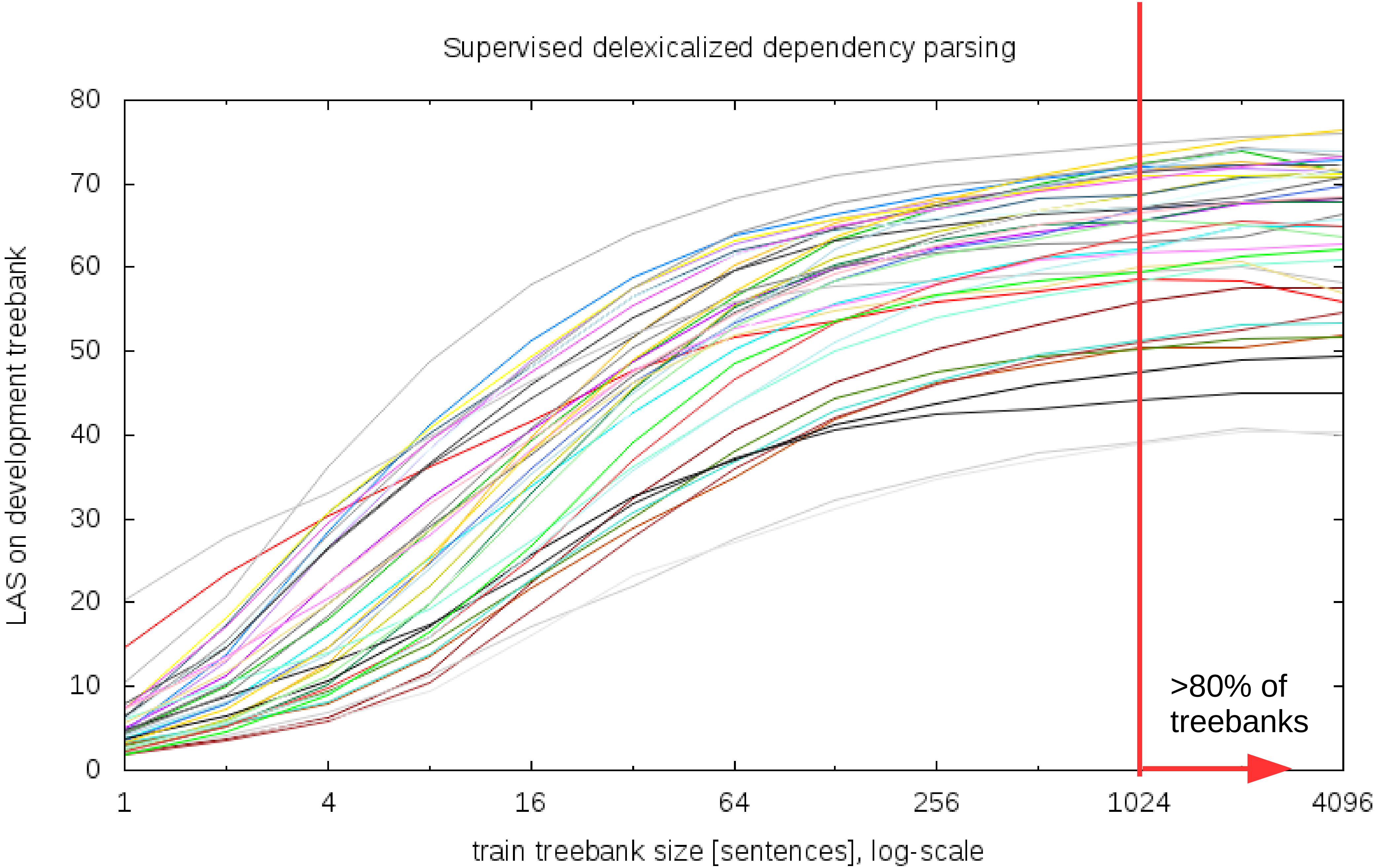
1. multilingual NMT/embeddings

- multilingual word embeddings
  - preliminary experiments (in 2016): worse than MT
- multilingual neural machine translation
  - have not tried, but looks promising, keen to try
  - zero-shot: jointly train A→B & B→C, translate A→C
    - results not terrific, but probably worth trying out
  - various problems to solve
    - constrain to 1:1 translation?
    - estimate word alignment from attention?
    - …not applicable to our setting out-of-the-box

2. looking at treebank size

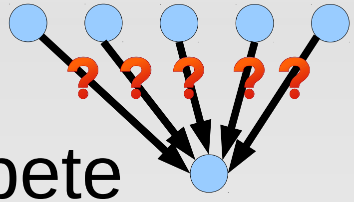


2. looking at treebank size



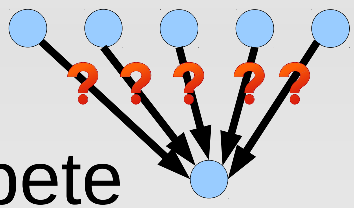
3. transition-based edge scores

- graph-based parser (MST) scores <u>all edges</u>
  - choose one head for each node
  - all N edges with the same dependent compete
    - their scores probably meaningfully comparable

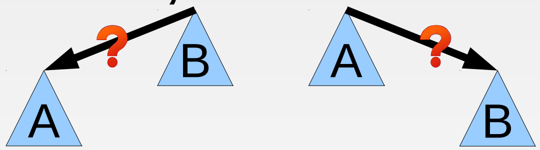


3. transition-based edge scores

- graph-based parser (MST) scores <u>all edges</u>
  - choose one head for each node
  - all N edges with the same dependent compete



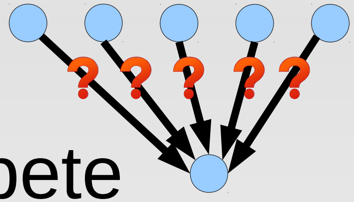
    - their scores probably meaningfully comparable
- transition-based parser (Malt) scores <u>operations</u>
  - choose 1 of 3 operations: left-arc, right-arc, shift
  - only 2 edges compete at each time



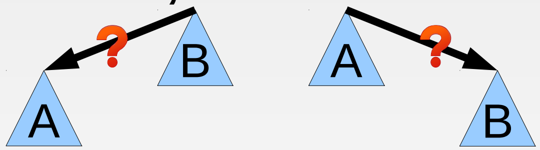
    - and they are just two directions of the same edge
    - scores probably not comparable across edges

3. transition-based edge scores

- graph-based parser (MST) scores all edges
  - choose one head for each node
  - all N edges with the same dependent compete
    - their scores probably meaningfully comparable



- transition-based parser (Malt) scores operations
  - choose 1 of 3 operations: left-arc, right-arc, shift
  - only 2 edges compete at each time
    - and they are just two directions of the same edge
    - scores probably not comparable across edges



  - possible path: beam search with global learning
    - scores entire transition sequences – not factored!

4. monolingual target texts

- source-target parallel data probably small
- larger monolingual target data may exist and may be exploited

4. monolingual target texts

- source-target parallel data probably small
- larger monolingual target data may exist and may be exploited
  - machine translation
    - train language model on larger data
    - expand parallel data via back-translation
      - translate target→source, add to para data
  - cross-lingual tagging
    - simple self-training: tag the data, retrain the tagger
  - cross-lingual parsing
    - pre-train word embeddings on larger data