Rudolf Rosa, Ondřej Dušek, David Mareček, Martin Popel
{rosa,odusek,marecek,popel}@ufal.mff.cuni.cz

Using Parallel Features in Parsing of Machine-Translated Sentences for Correction of Grammatical Errors

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

SSST, Jeju, 12th July 2012

Parsing of SMT Outputs

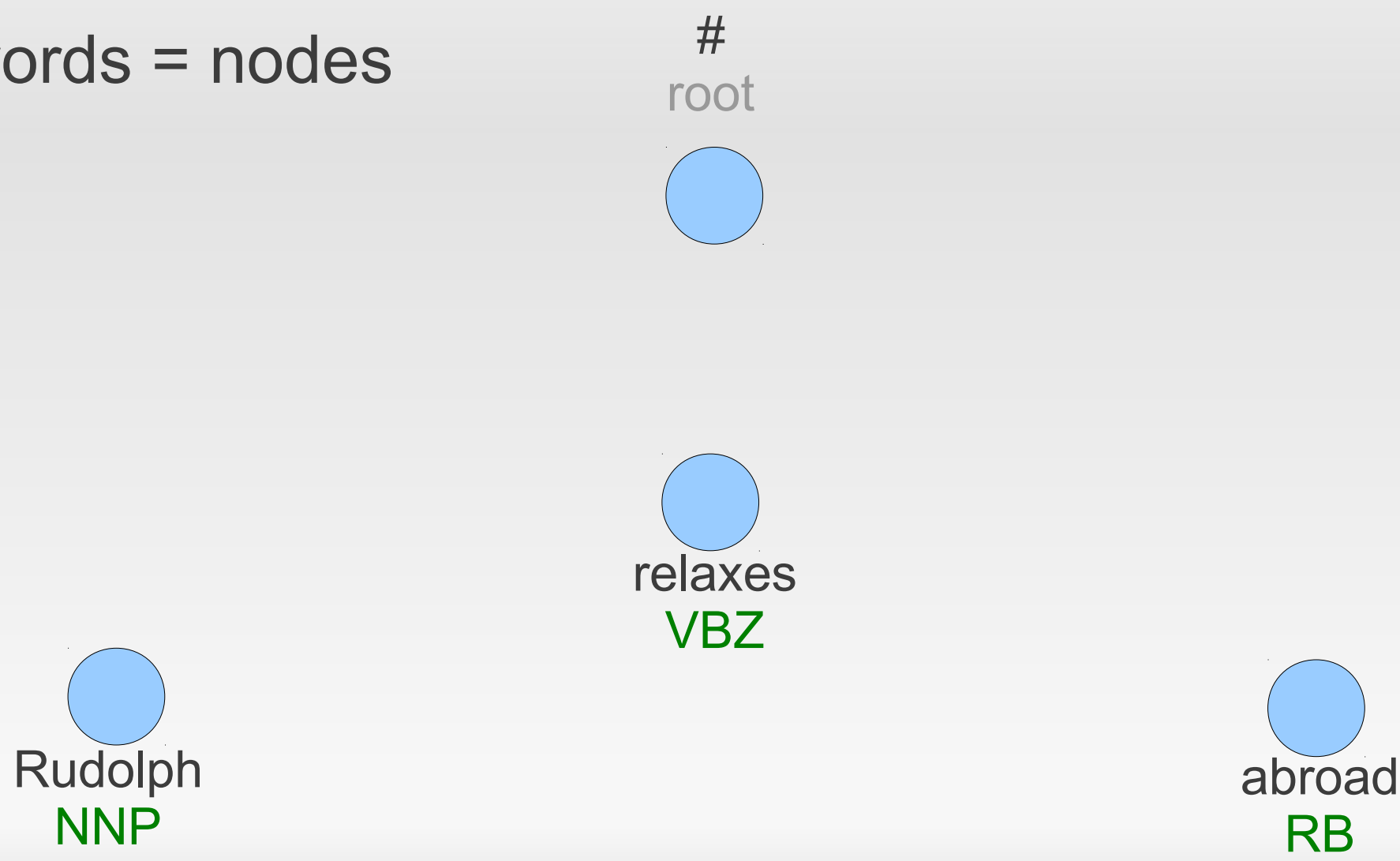
- can be useful in many applications
  - automatic classification of translation errors
  - **automatic correction of translation errors (Depfix)**
  - confidence estimation, multilingual question answering...
- ✔ we have the source sentence available
  - Can we use it to help parsing?
- ✗ SMT outputs noisy (errors in fluency, grammar...)
  - parsers trained on gold standard treebanks
  - Can we adapt parser to noisy sentences?

MST Parser

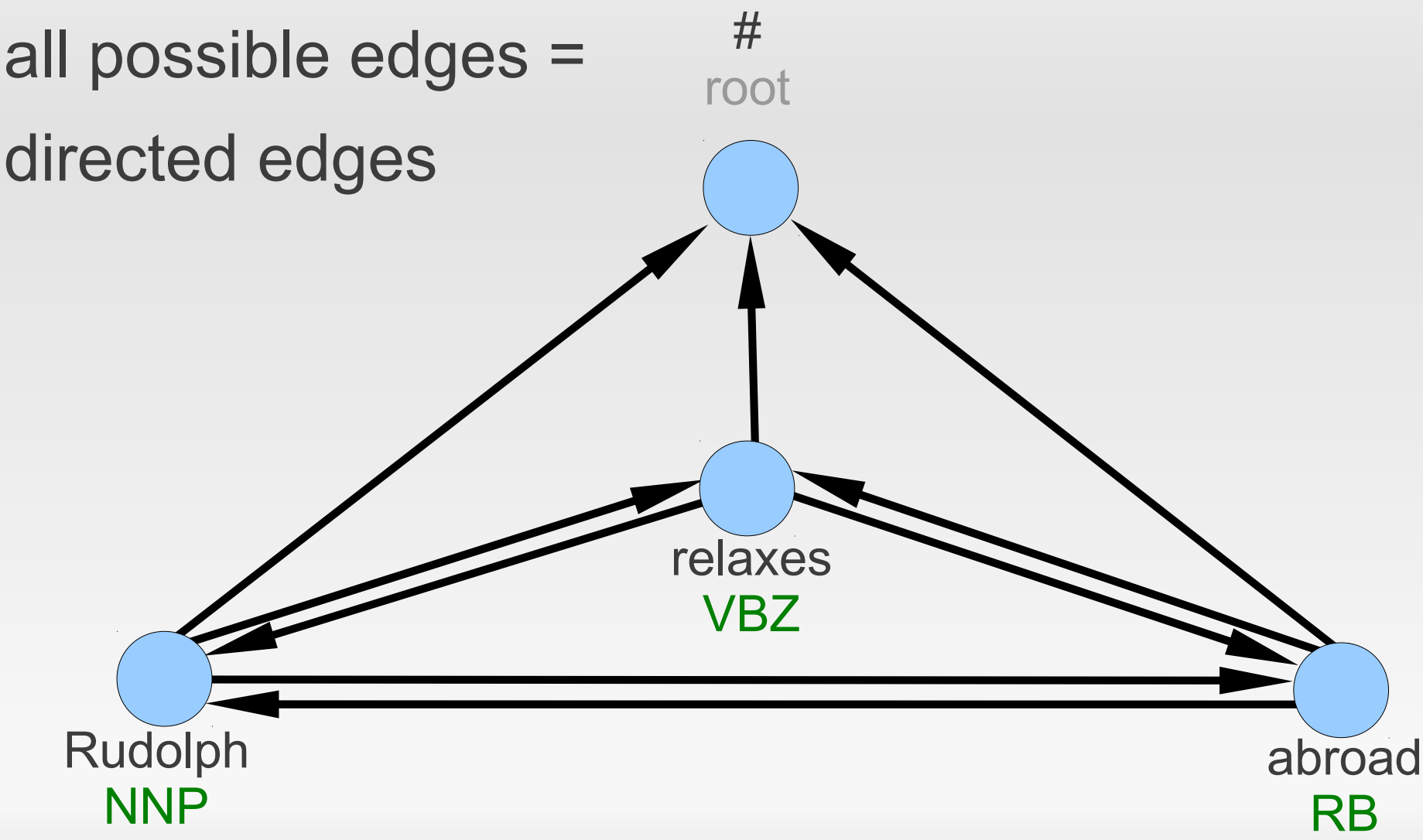
- Maximum Spanning Tree dependency parser
- by Ryan McDonald

(1) Words and Tags

words = nodes



(2) (Nearly) Complete Graph

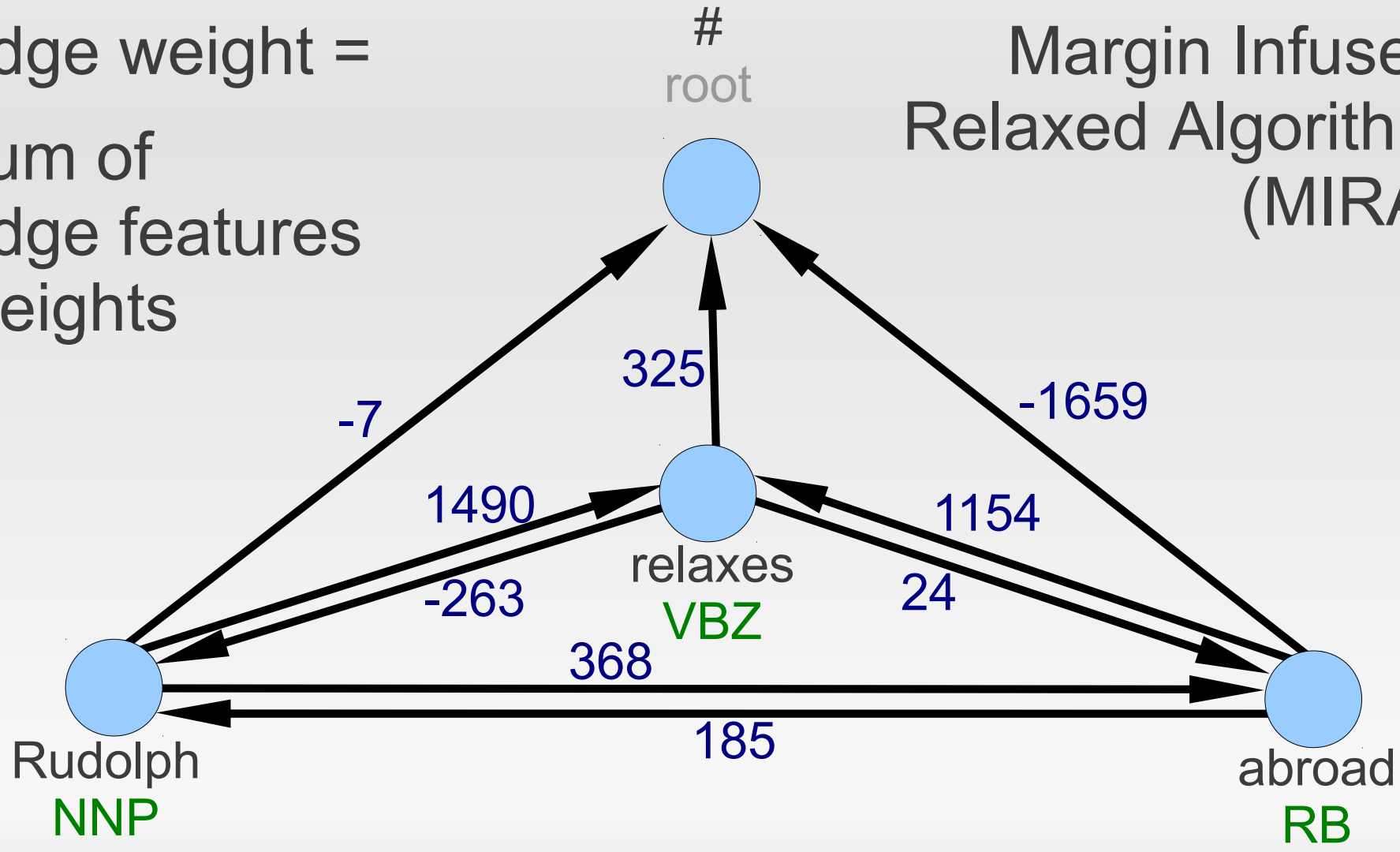


(3) Assign Edge Weights

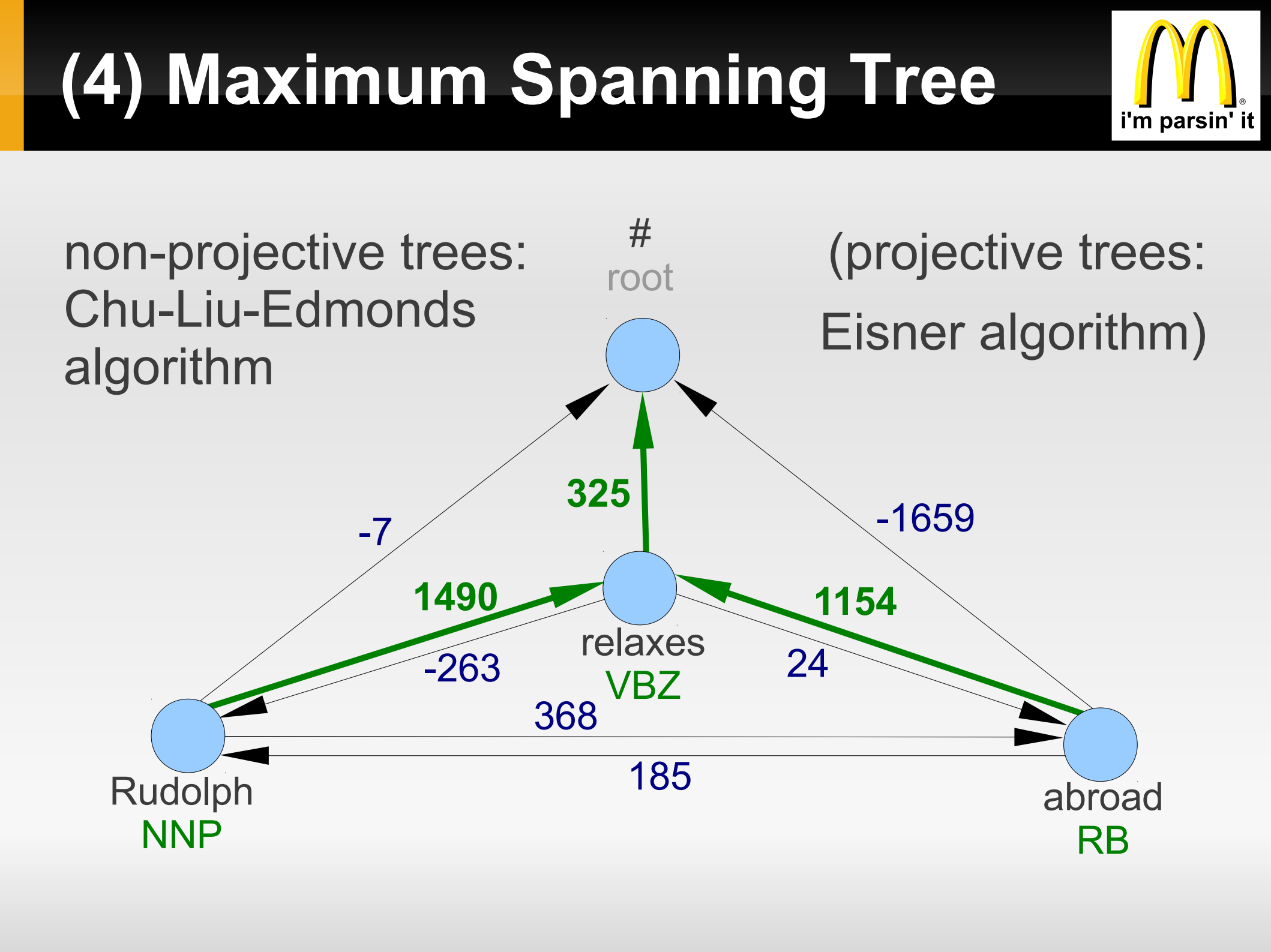
edge weight =

sum of
edge features
weights

Margin Infused
Relaxed Algorithm
(MIRA)



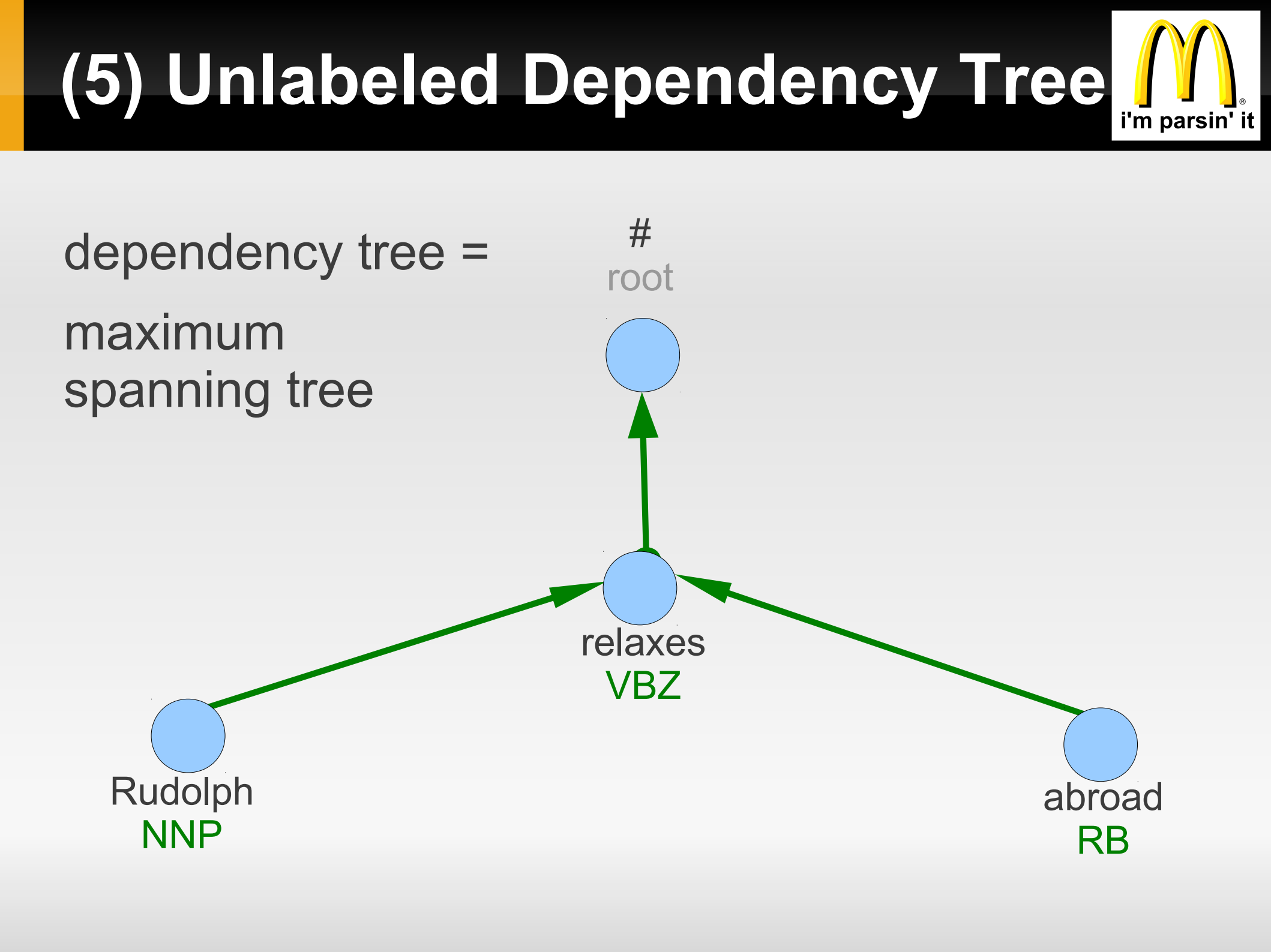
(4) Maximum Spanning Tree



non-projective trees:
Chu-Liu-Edmonds
algorithm

(projective trees:
Eisner algorithm)

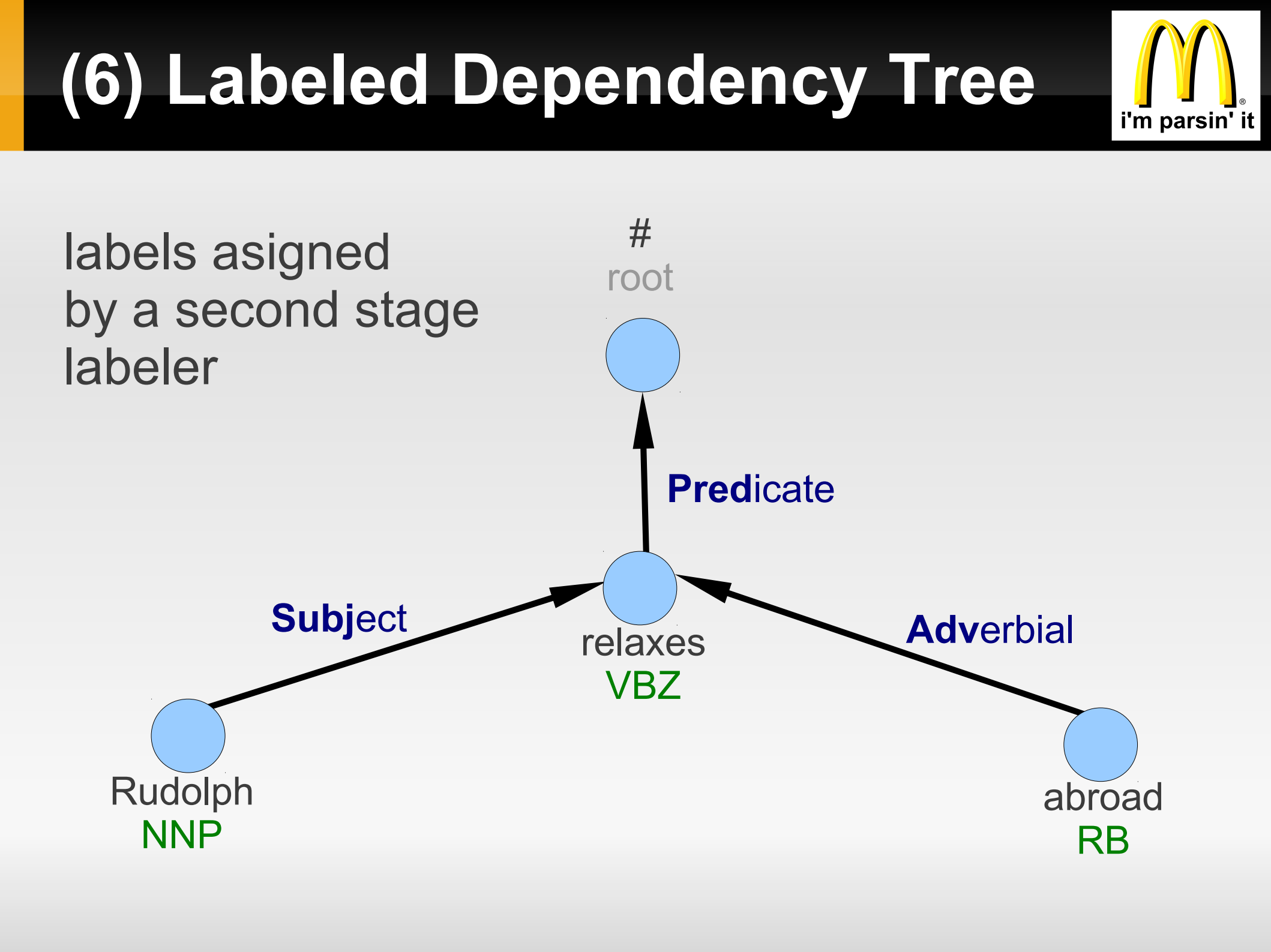
(5) Unlabeled Dependency Tree



dependency tree =

maximum spanning tree

(6) Labeled Dependency Tree



labels asigned
by a second stage
labeler

RUR Parser

- reimplementation of MST Parser
  - (so far only) first-order, non-projective
- adapted for SMT outputs parsing
  - parallel features
  - ”worsening” the training treebank

English-to-Czech SMT

- Czech language
  - highly flective
    - 4 genders, 2 numbers, 7 cases, 3 persons...
    - Czech grammar requires agreement in related words
  - word order relatively free: word order errors not crucial
- Phrase-Based SMT often makes inflection errors:
  - ➔ Rudolph's car is black.
  - ✗ Rudolfov**a**/***fem*** auto/***neut*** je černý/***masc***.
  - ✔ Rudolfov**o**/***neut*** auto/***neut*** je černé/***neut***.

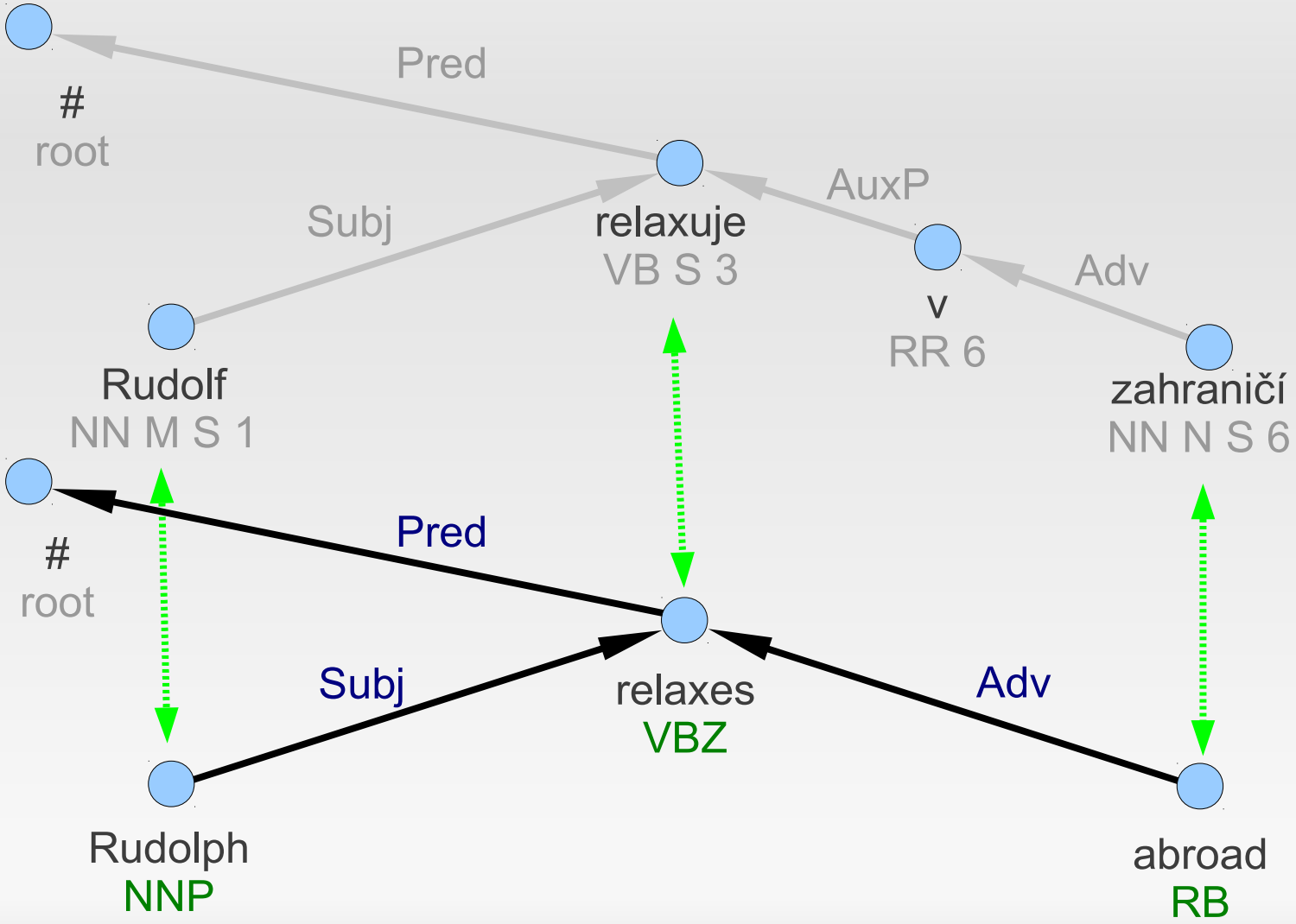
Parser Training Data

- Prague Czech-English Dependency Treebank
  - parallel treebank
  - 50k sentences, 1.2M words
  - morphological tags, surface syntax, deep syntax
  - word alignment

Parallel Features

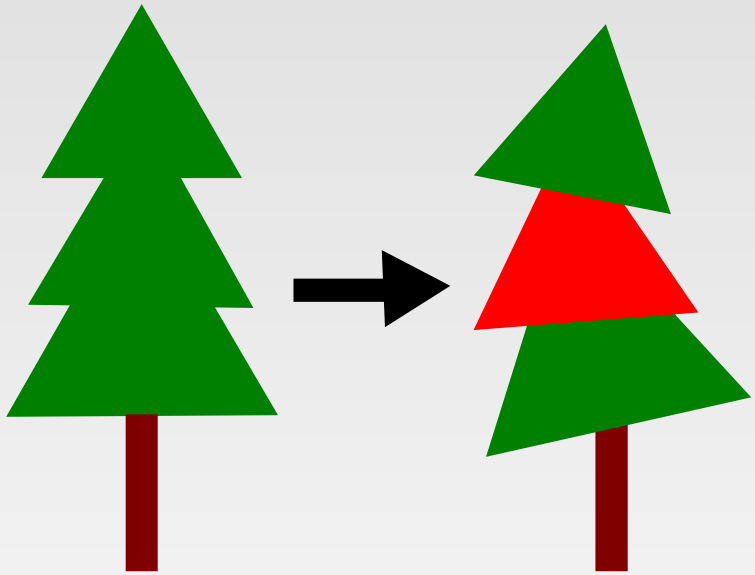
- word alignment (using GIZA++)
- additional features (if aligned node exists):
  - aligned tag (NNS, VBD...)
  - aligned dependency label (Subject, Attribute...)
  - aligned edge existence (0/1)

Parallel Features Example

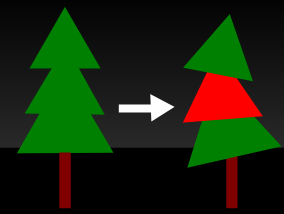


Worsening the Treebank

- treebank used for training contains correct sentences
- SMT output is noisy
  - grammatical errors
  - incorrect word order
  - missing/superfluous words
  - …
- let's introduce similar errors into the treebank!
  - so far, we have only tried inflection errors

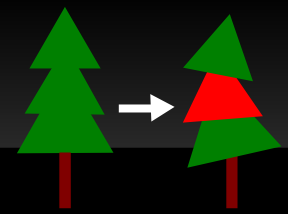


Worsen (1): Apply SMT



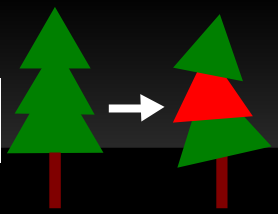
- translate English side of PCEDT to Czech
  - by an SMT system (we used Moses)
- now we have (e.g.):
  - Gold English
    - Rudolph's car is black.
  - Gold Czech
    - Rudolfovo$_{NEUT}$ auto$_{NEUT}$ je černé$_{NEUT}$.
  - SMT Czech
    - Rudolfova$_{FEM}$ auto$_{NEUT}$ je černý$_{MASC}$.

Worsen (2): Align SMT to Gold



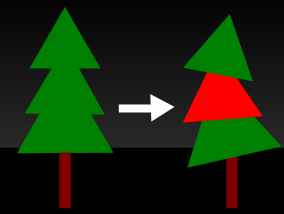
- align SMT Czech to Gold Czech
- Monolingual Greedy Aligner
  - alignment link score = linear combination of:
    - similarity of word forms (or lemmas)
    - similarity of morphological tags (fine-grained)
    - similarity of positions in the sentence
    - indication whether preceding/following words aligned
  - repeat: align best scoring pair until below threshold
  - no training: weights and threshold set manually

Worsen (3): Create Error Model



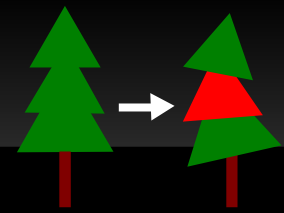
- for each tag:
  - estimate probabilities of SMT system using an incorrect tag instead of the correct tag (Maximum Likelihood Estimate)
- Czech tagset: fine-grained morphological tags
  - part-of-speech, gender, number, case, person, tense, voice...
  - 1500 different tags in training data

Worsen (3): Error Model



- Adjective, Masculine, Plural, Instrumental case (`AAMP7`), e.g. *lingvistickými* (linguistic)
  - **0.2** Adjective, Masculine, Singular, Nominative case
    - e.g. *lingvistický*
  - **0.1** Adjective, Masculine, Plural, Nominative case
    - e.g. *lingvističtí*
  - **0.1** Adjective, Neuter, Singular, Accusative case
    - e.g. *lingvistické*
- … altogether 2000 such change rules

Worsen (4): Apply Error Model



- take Gold Czech
- for each word:
  - assign a new tag randomly sampled according to Tag Error Model
  - generate a new word form
    - rule-based generator, generates even unseen forms
    - new_form = generate_form(lemma, tag) || old_form
- → get Worsened Czech
- use resulting *Gold English-Worsened Czech* parallel treebank to train the parser

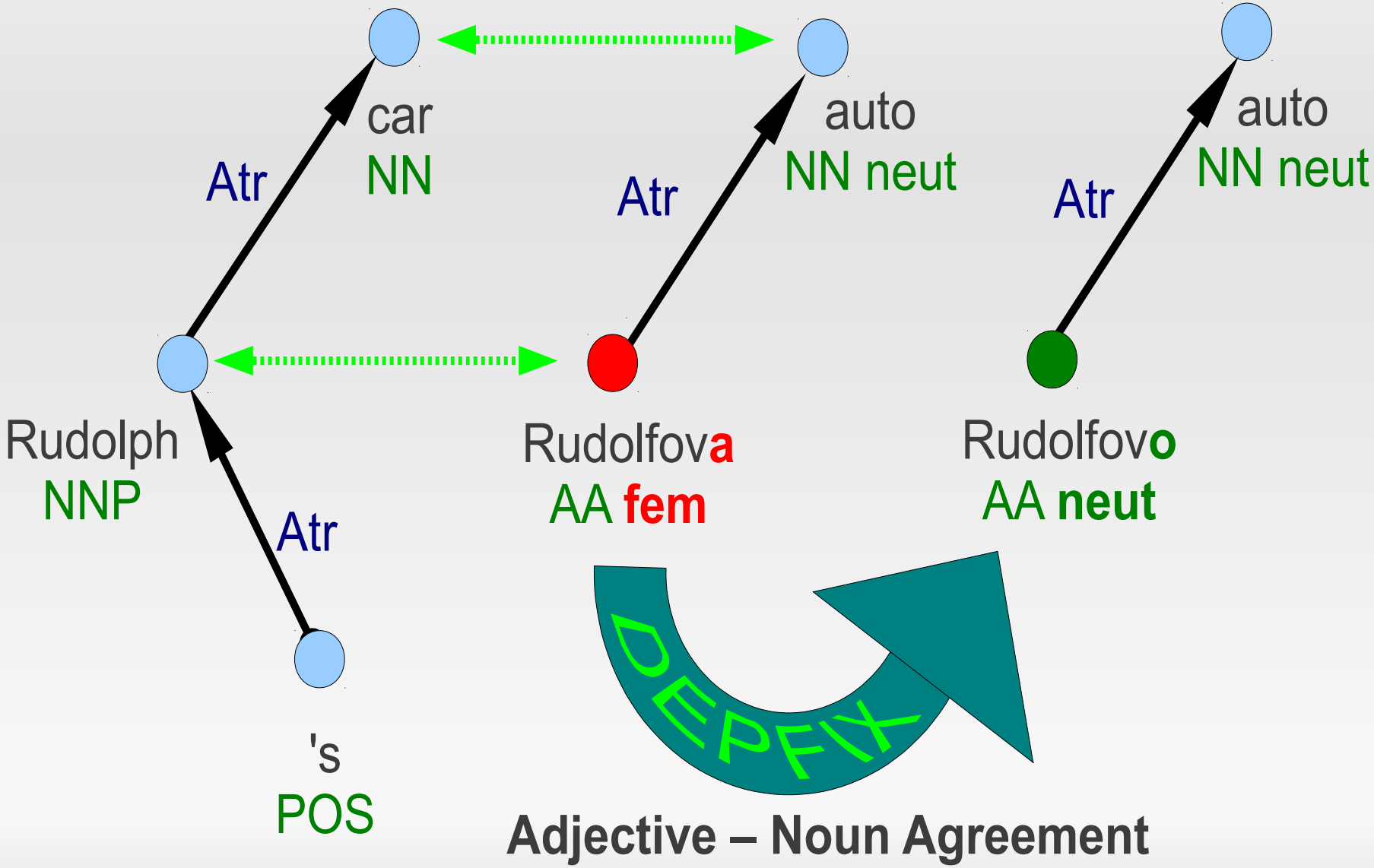
Direct Evaluation by Inspection

- manual inspection of several parse trees
  - comparing baseline and adapted parser ouputs
- examples of improvements:
  - subject identification even if not in nominative case
  - adjective-noun dependence identification even if agreement violated (gender, number, case)
- hard to do reliably
  - trying to find a correct parse tree for an (often) incorrect sentence – not well defined

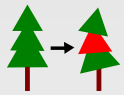
Indirect Evaluation: in Depfix

- rule-based grammar correction of SMT outputs
- input = aligned, tagged and **parsed** sentences:
  - target (Czech) sentence – to be corrected
  - source (English) sentence – additional information
- applies 20 correction rules:
  - noun – adjective agreement (gender, number, case)
  - subject – predicate agreement (gender, number)
  - preposition – noun agreement (case)
  - …

Depfix: Rudolph's Car



Indirect Evaluation Results

- differences in Depfix corrections evaluated by humans: better / worse / indefinite
- three different parsers
  - RUR + parallel features + worsened treebank 
  - 

 – original McDonald's MST Parser
  - RUR – our baseline setup

	RUR + parallel features + worsened treebank		
	better	worse	indefinite
	**51%**	30%	18%
RUR	**54%**	28%	18%

Conclusion

- SMT outputs often hard to parse
- RUR parser – adapted to parsing SMT outputs
  - parallel features (tag, dep. label, edge existence)
  - worsening the training treebank (tag error model)
- outputs of English-to-Czech translation
- evaluated in Depfix
  - SMT errors correction system

Future Work

- more sophisticated parallel features
- more experiments on worsening
- more languages

- parallel tagging

Thank you for your attention

For this presentation and other information, visit:

http://ufal.mff.cuni.cz/~rosa/depfix/

Rudolf Rosa, Ondřej Dušek, David Mareček, Martin Popel
{rosa,odusek,marecek,popel}@ufal.mff.cuni.cz

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