Rudolf Rosa
rosa@ufal.mff.cuni.cz

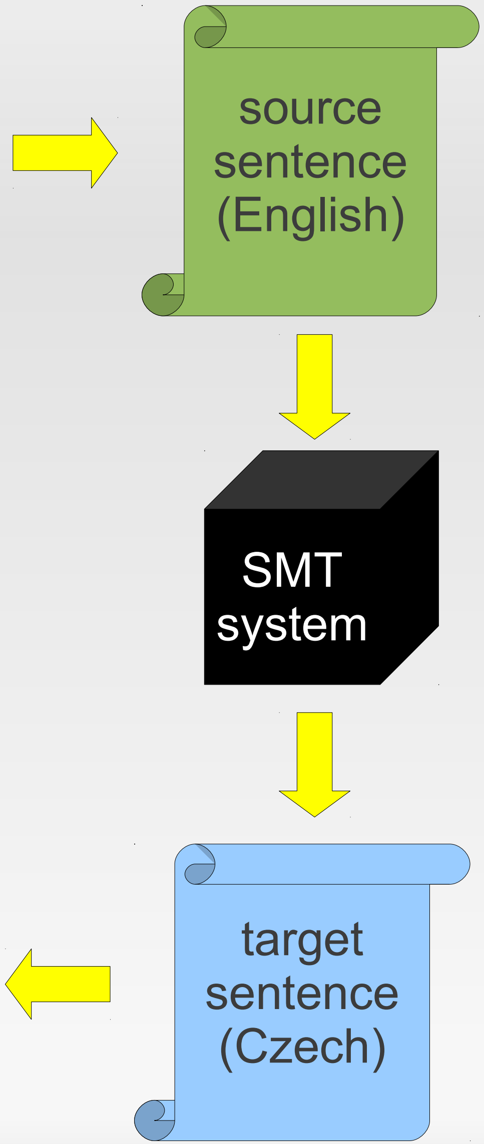
Depfix:

Automatic post-editing of phrase-based machine translation outputs

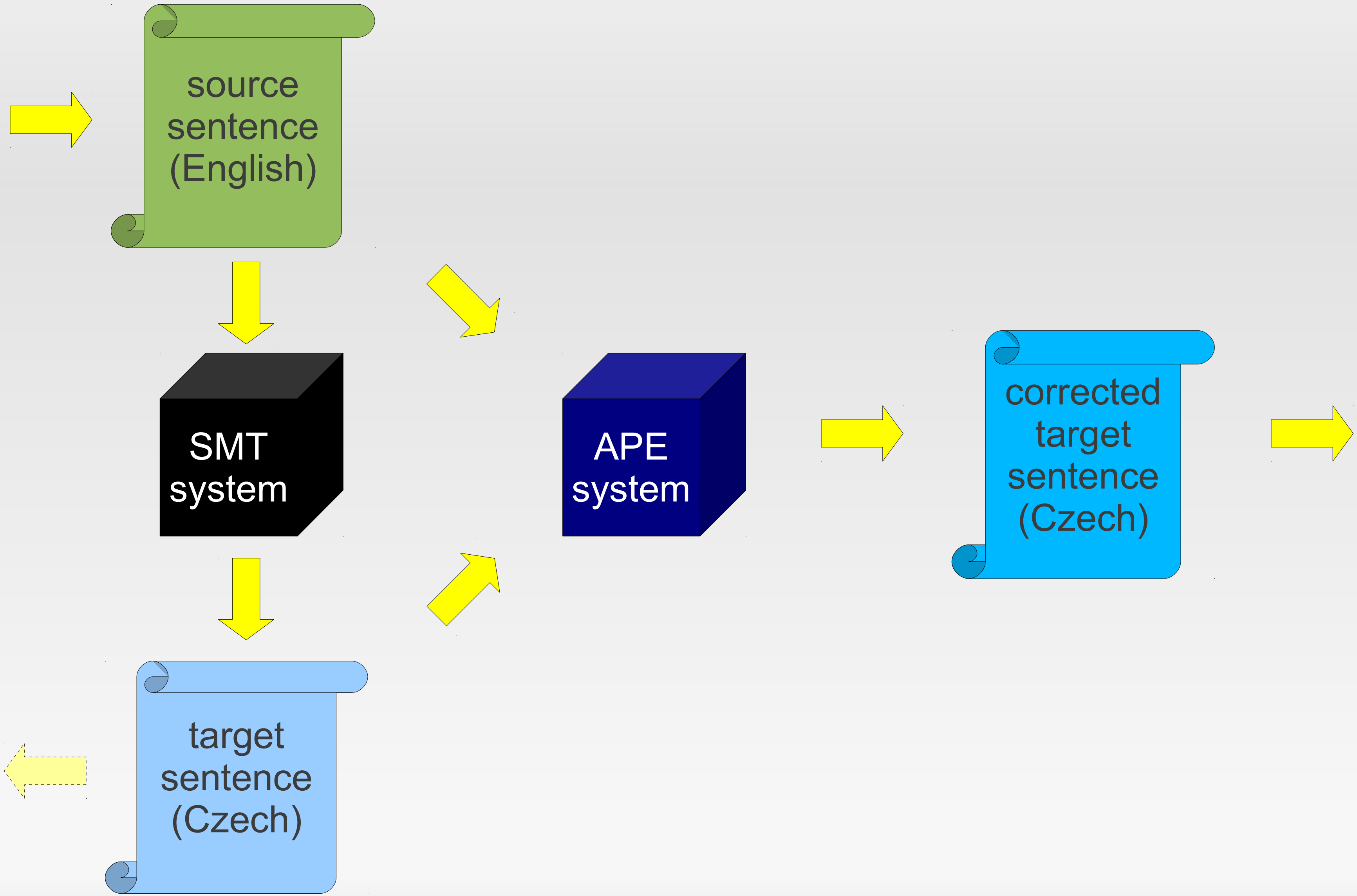
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Edinburgh SMT group meeting, 29th January 2014

Statistical Machine Translation



Automatic Post-editing of SMT



Depfix

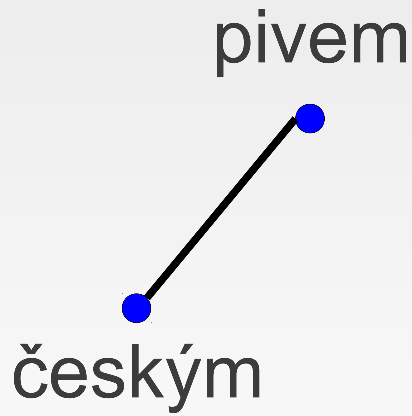
- input: target sentence (+ source sentence)

1. analysis
   - tokenization, lemmatization, tagging, word-alignment, parsing, deep-syntax induction
2. correction
   - a set of rules, e.g. noun-adjective agreement
3. generation
   - morphological generator

- output: corrected target sentence

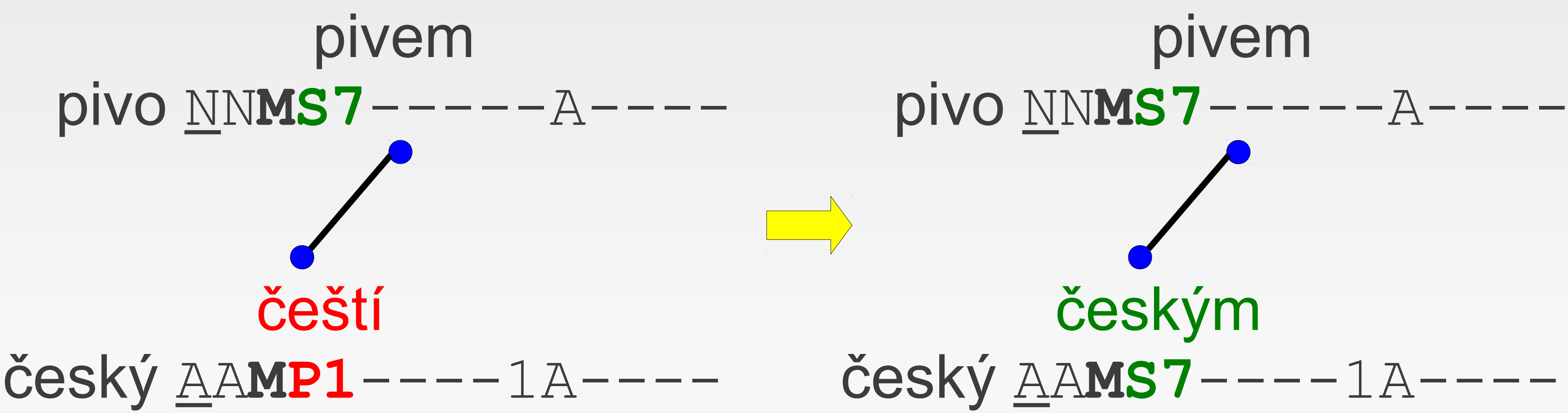
Analysis

- most important:
  - **lemma** + Czech fine-grained morphological **tag**
    - gender, number, case, person, tense, negation...
    - e.g. *pivem* (*beer* in instrumentative case)
    - → *pivo* `NNNS7-----A----`
  - dependency **tree**
    - tells us e.g. that the modifier *českým* (Czech) belongs to *pivem* (beer)
    - we use **MST parser adapted for parsing SMT outputs**



Correction

- usually edge-local
- morphological agreement of noun and adjective:
  - set **gender**, **number** and **case** of the adjective to **gender**, **number** and **case** of the noun



Generation

- morphological generator
  - lemma & tag → word form
  - e.g. *český* `AAMS7----1A----` → *českým*
- Czech morphology is far from trivial
  - 2 numbers, 4 genders, 7 cases, various paradigms...
  - homonymous forms
    - *piva* = sg gen / pl nom / pl acc / pl voc
  - variants
    - sg loc = *pivu / pivě*

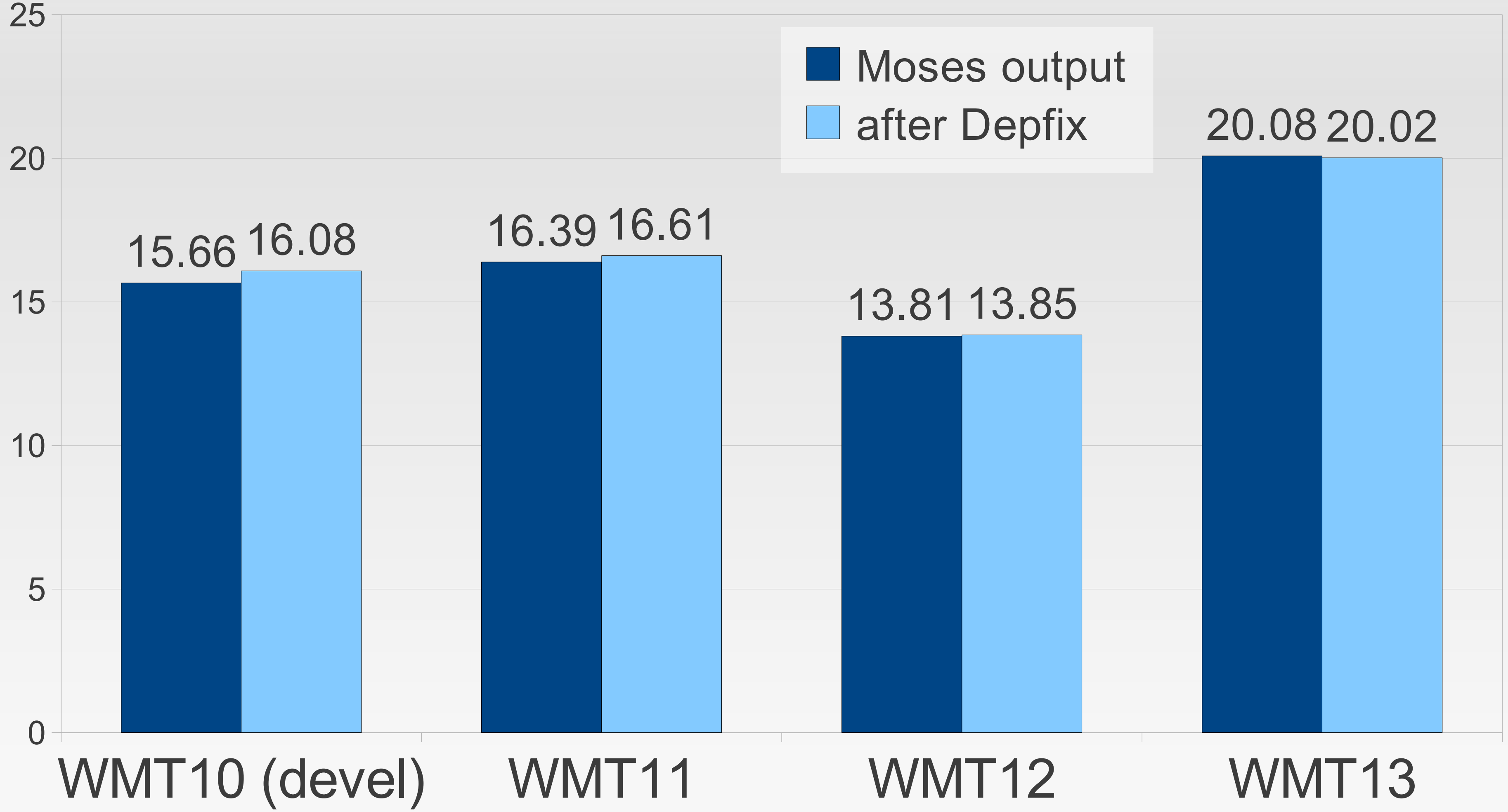
Correction types (I)

- agreement
  - preposition – noun (case)
  - noun – adjective (gender, number, case)
  - subject – predicate (gender, number, person)
  - antecedent – relative pronoun (gender, number, case)
- minor errors
  - projection of tokenization
  - source-aware truecasing
  - vocalization of prepositions

Correction types (II)

- transfer of meaning to morphology
  - translation of possessives and “of” (genitive)
  - translation of passive voice and “by” (instrumentative)
  - subject (nominative)
  - verb tense
  - negation
- coarse translation of missing items
  - missing reflexive verbs
- analysis corrections (alignment, tags, trees)

Automatic evaluation (BLEU)



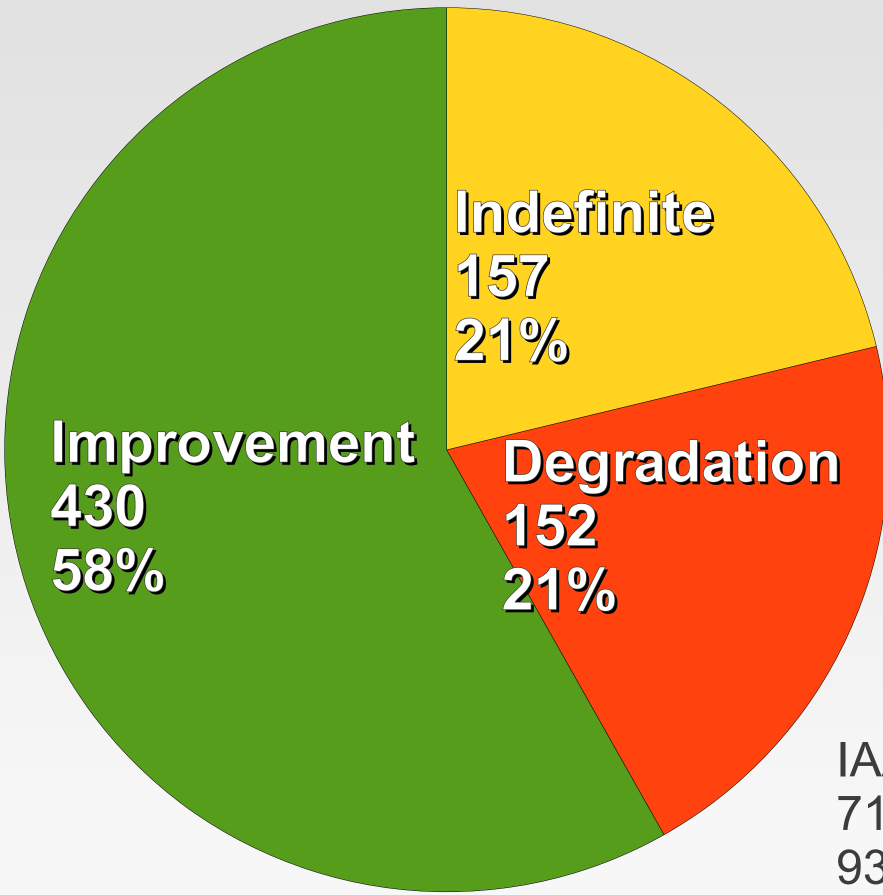
Automatic evaluation (Δ BLEU)

data/system	CU Bojar	CU TectoMT	CU Zeman	UEdin
WMT10 (dev)	+0.33	-0.07	+0.61	+0.78
WMT11	+0.47	-0.10	+0.73	+0.64
WMT12	+0.07	-0.02	+0.34	+0.23

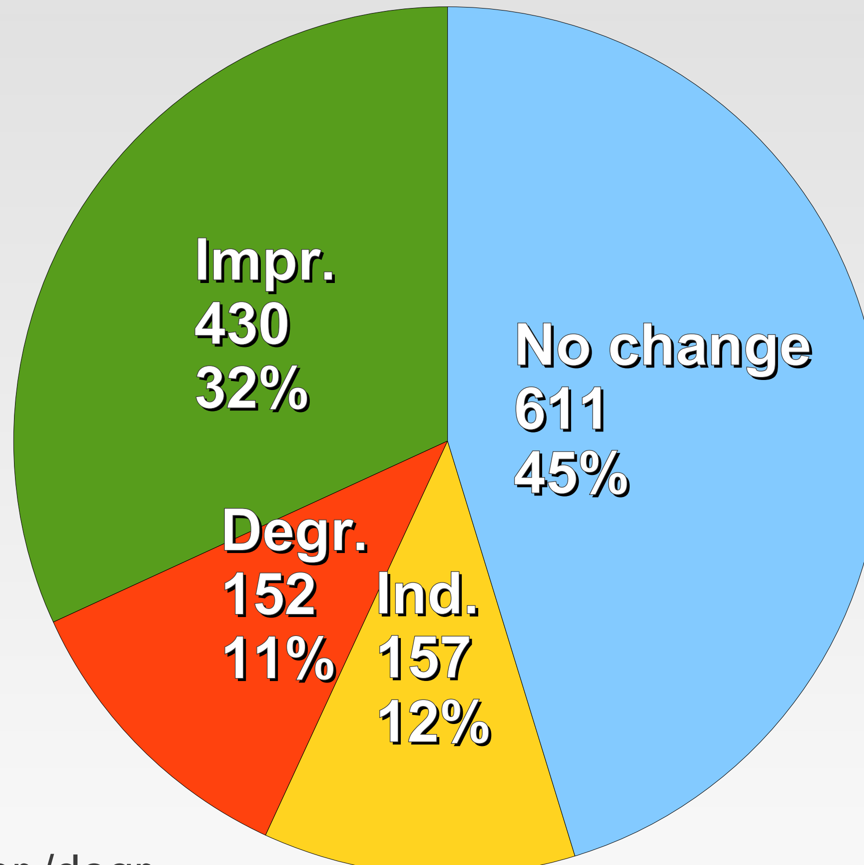
data/system	SFU	EuroTrans	Bing	Google Tr.
WMT10 (dev)	**+1.05**	+0.35	+0.78	+0.59
WMT11	–	+0.21	–	+0.23
WMT12	+0.41	+0.15	+0.37	0.00

Manual evaluation (WMT12)

- changed sentences



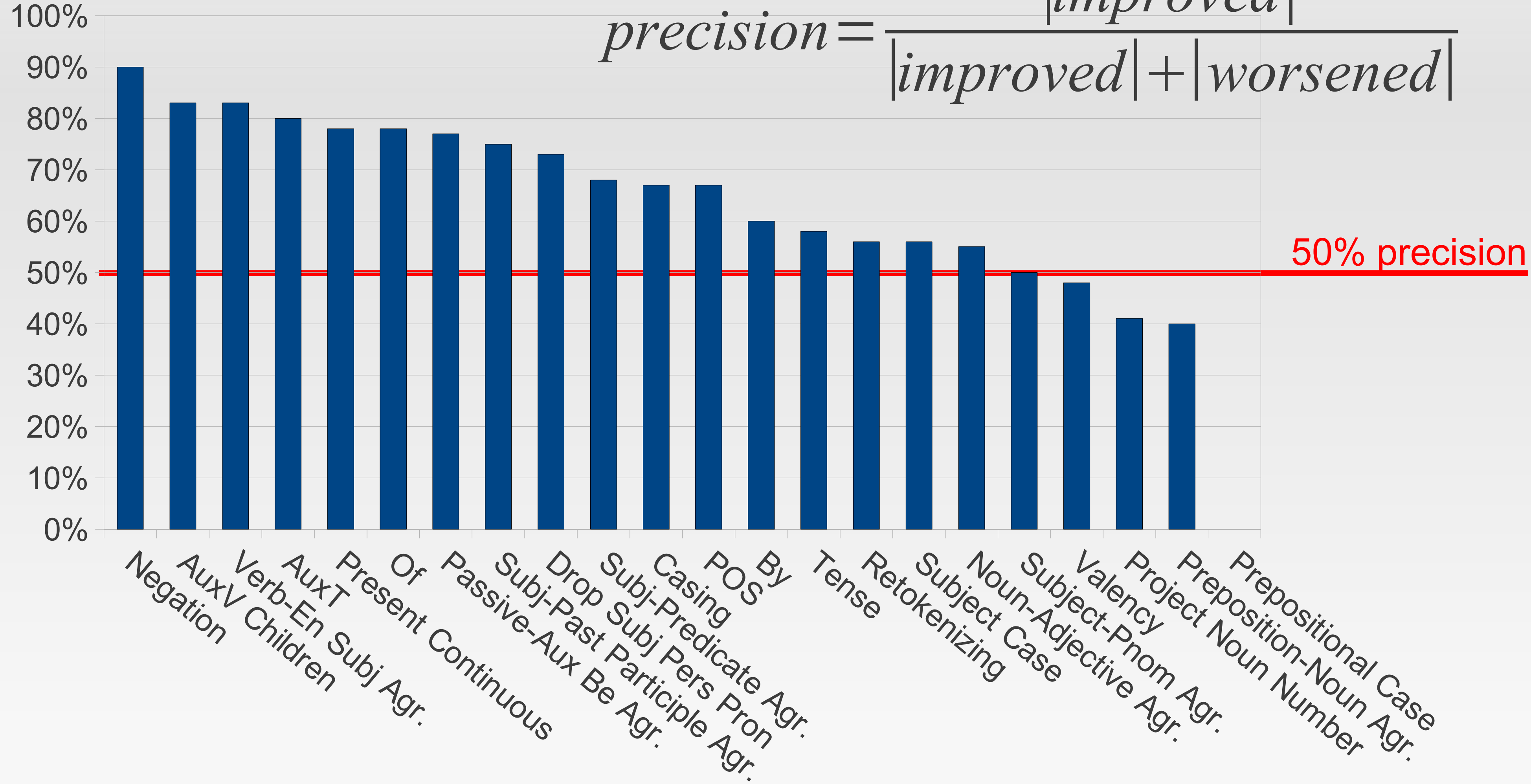
- all sentences



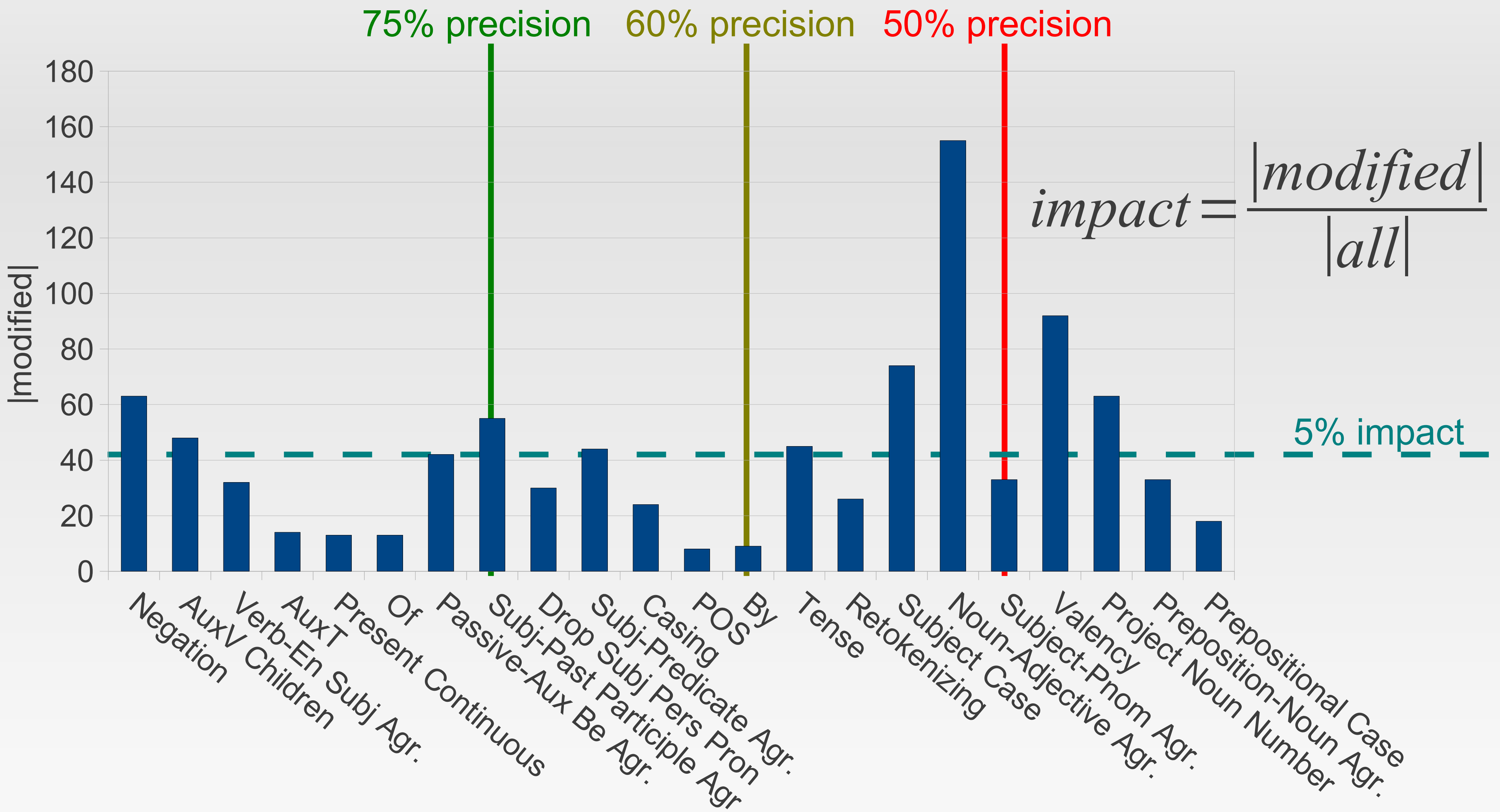
IAA:
71% all
93% impr./degr.

Precision of rules (part of WMT12)

$$precision = \frac{|improved|}{|improved| + |worsened|}$$



Impact of rules (part of WMT12)



$$impact = \frac{|modified|}{|all|}$$

Current & future work

- if we can write the rules manually...
  - …can we also machine-learn them?
- currently running experiments
  - predict: gender, number, case (of modifier)
  - data: parallel corpus & its translation by Moses
  - decision trees / maximum entropy classifier
  - features (modifier, head & their source counterparts)
    - tag (split by category), dependency relation label, edge direction, number of modifiers, lemma
  - preliminary positive results

Delving deeper...

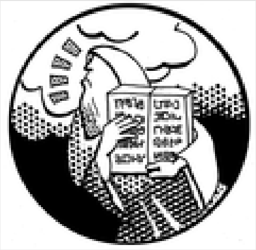
- more corrections
  - an example of a cascade of corrections
  - correction of negation
  - correction of verb tense translation
- MSTperl parser
  - reimplementation of MST parser
  - adapted for parsing SMT outputs

Delving deeper...

- more corrections
  - **an example of a cascade of corrections**
  - correction of negation
  - correction of verb tense translation
- MSTperl parser
  - reimplementation of MST parser
  - adapted for parsing SMT outputs

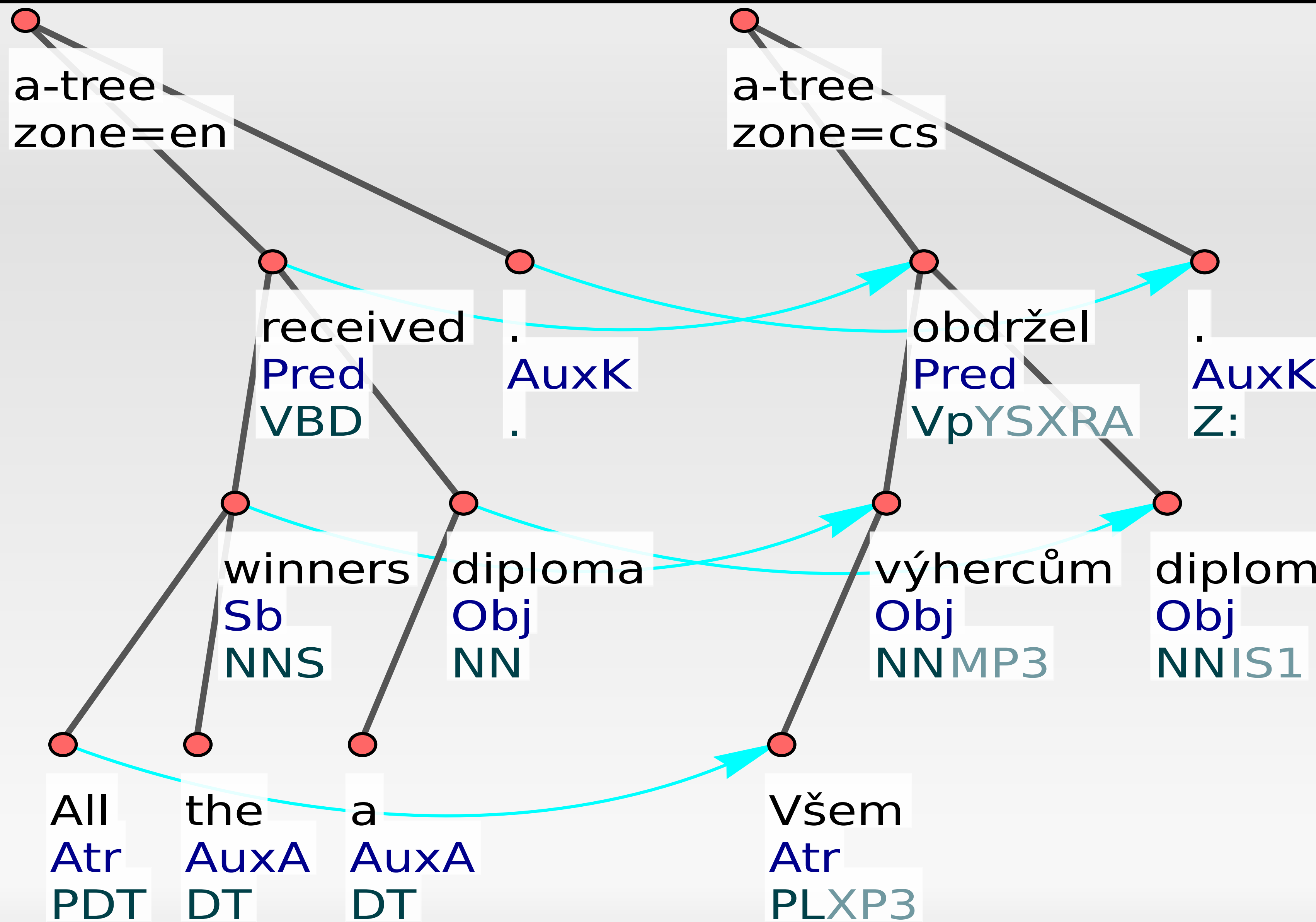
A cascade of corrections

- Source:
  - *All the winners received a diploma.*
- Moses:
  - *Všem výhercům obdržel diplom.*
  - *To all the winners he received a diploma.*

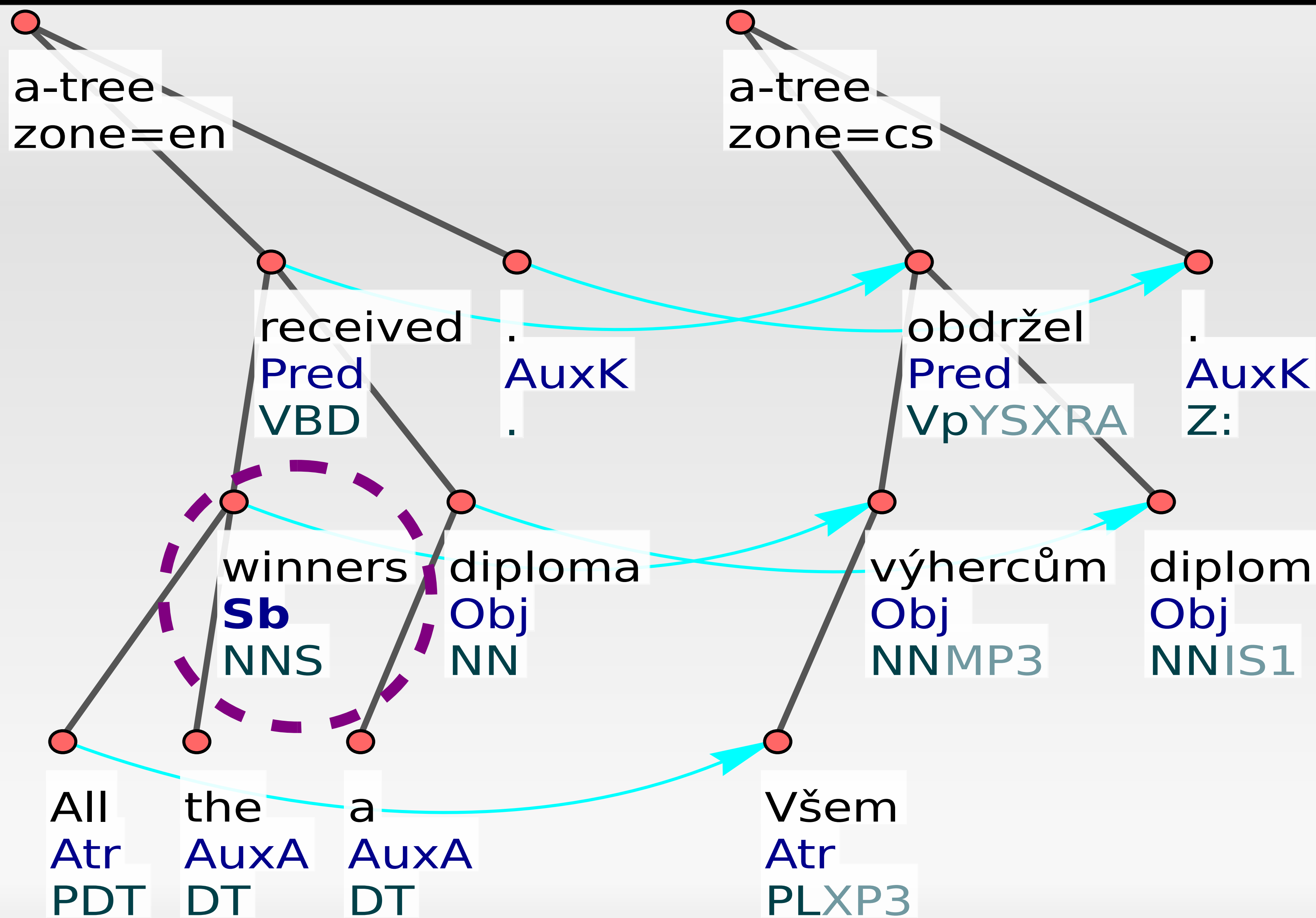


- Depfix:
  - *Všichni výherci obdrželi diplom.*
  - *All the winners received a diploma.*

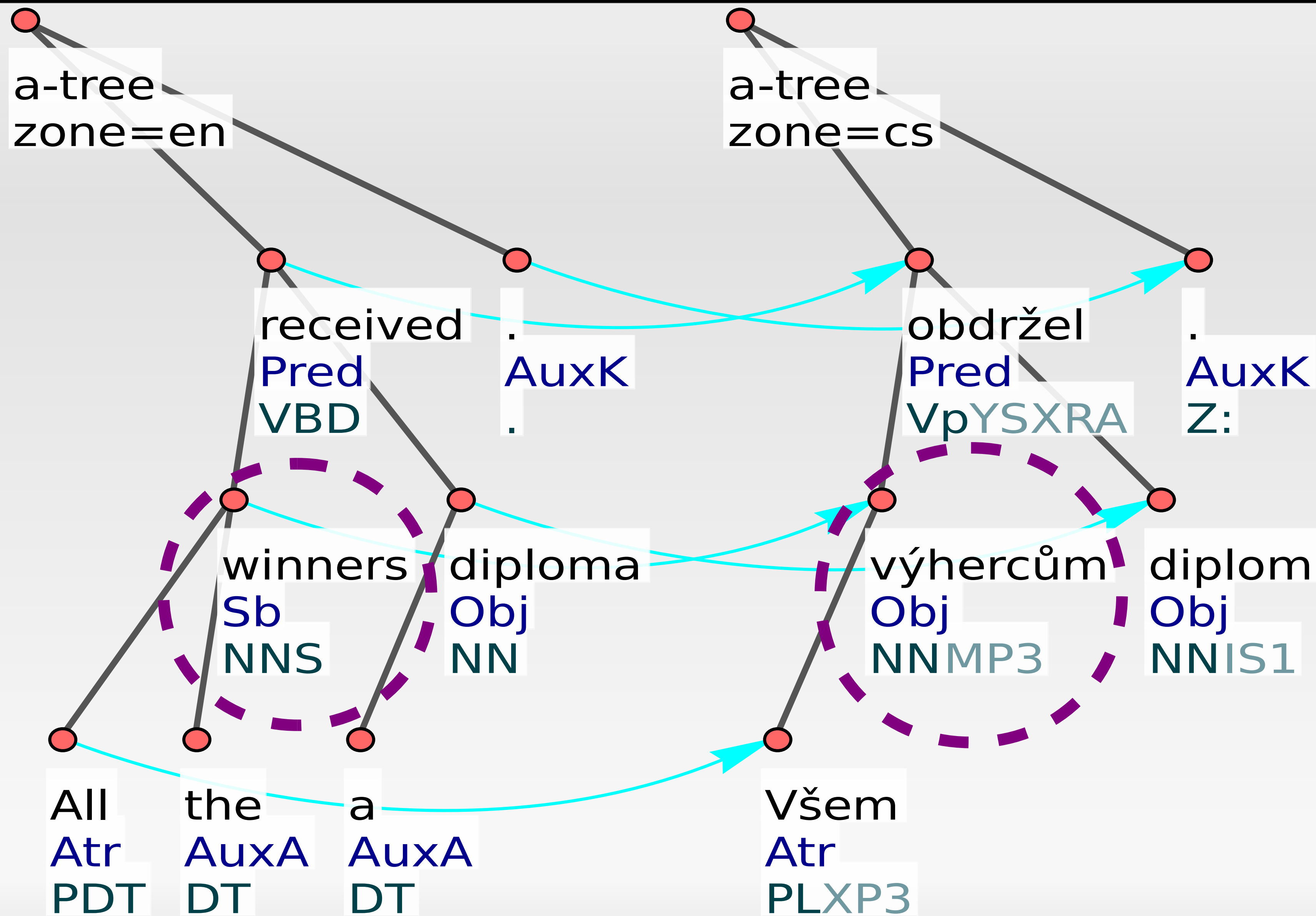
Všem výhercům obdržel diplom.



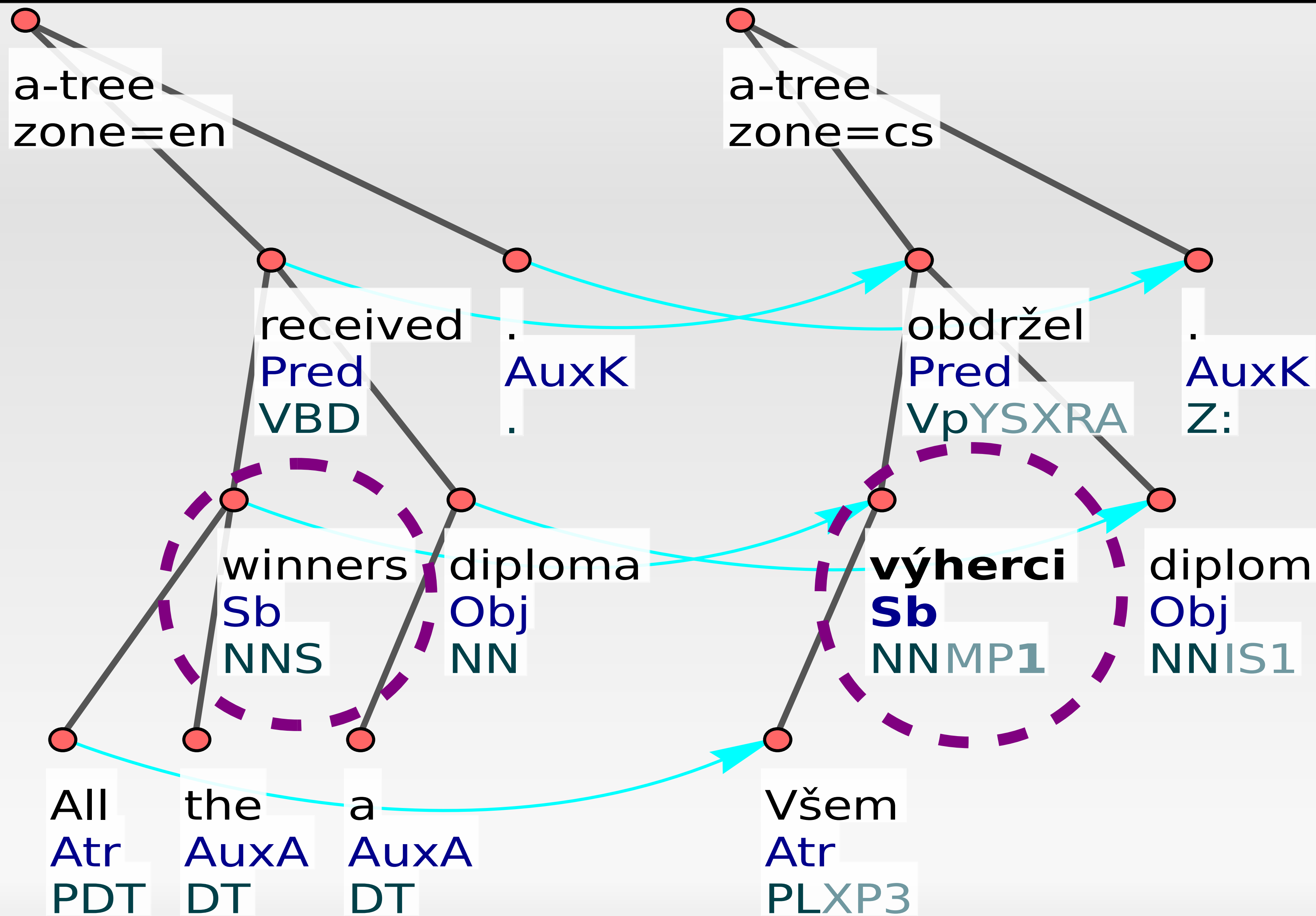
Transfer of meaning: subject



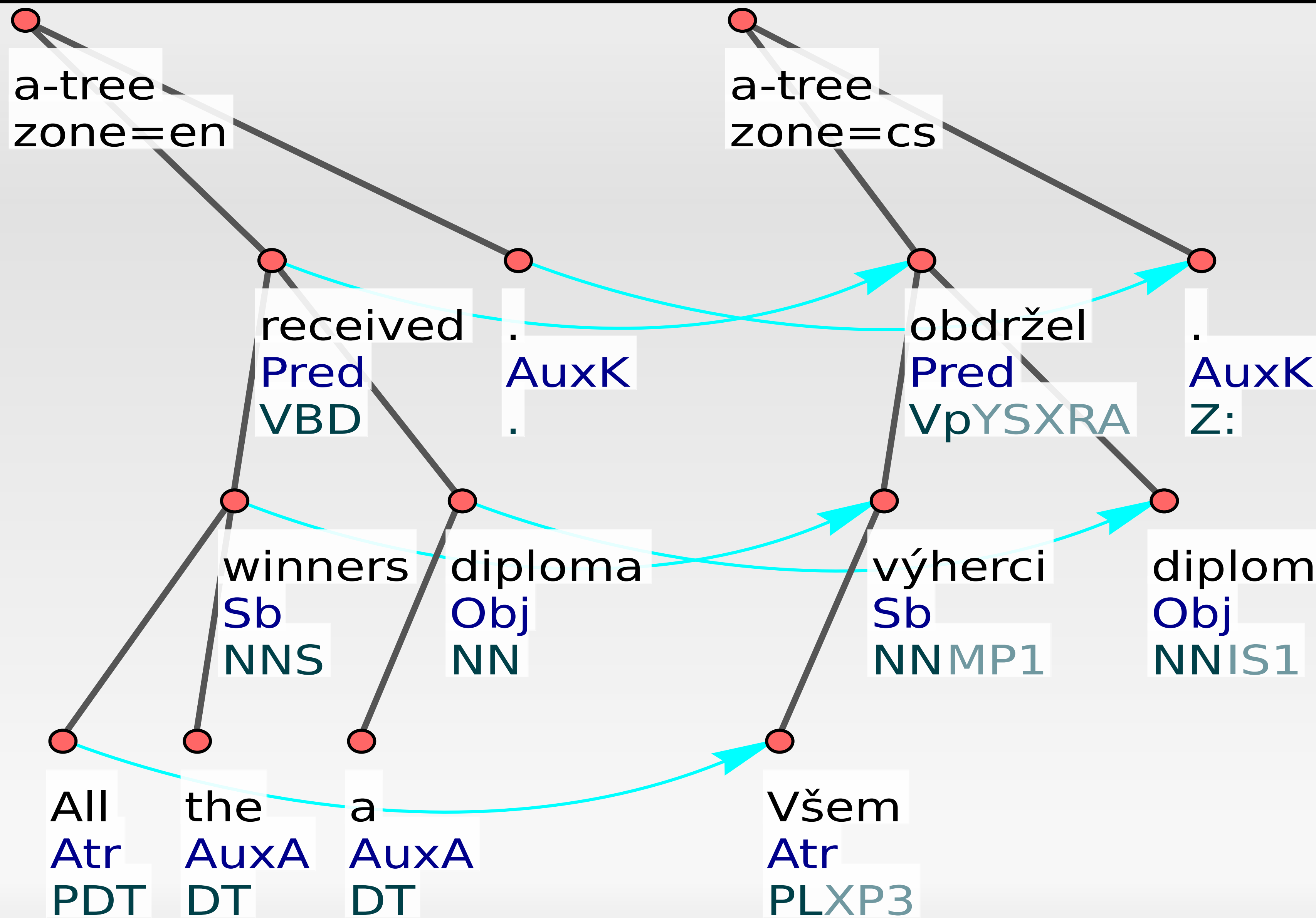
Transfer of meaning: subject



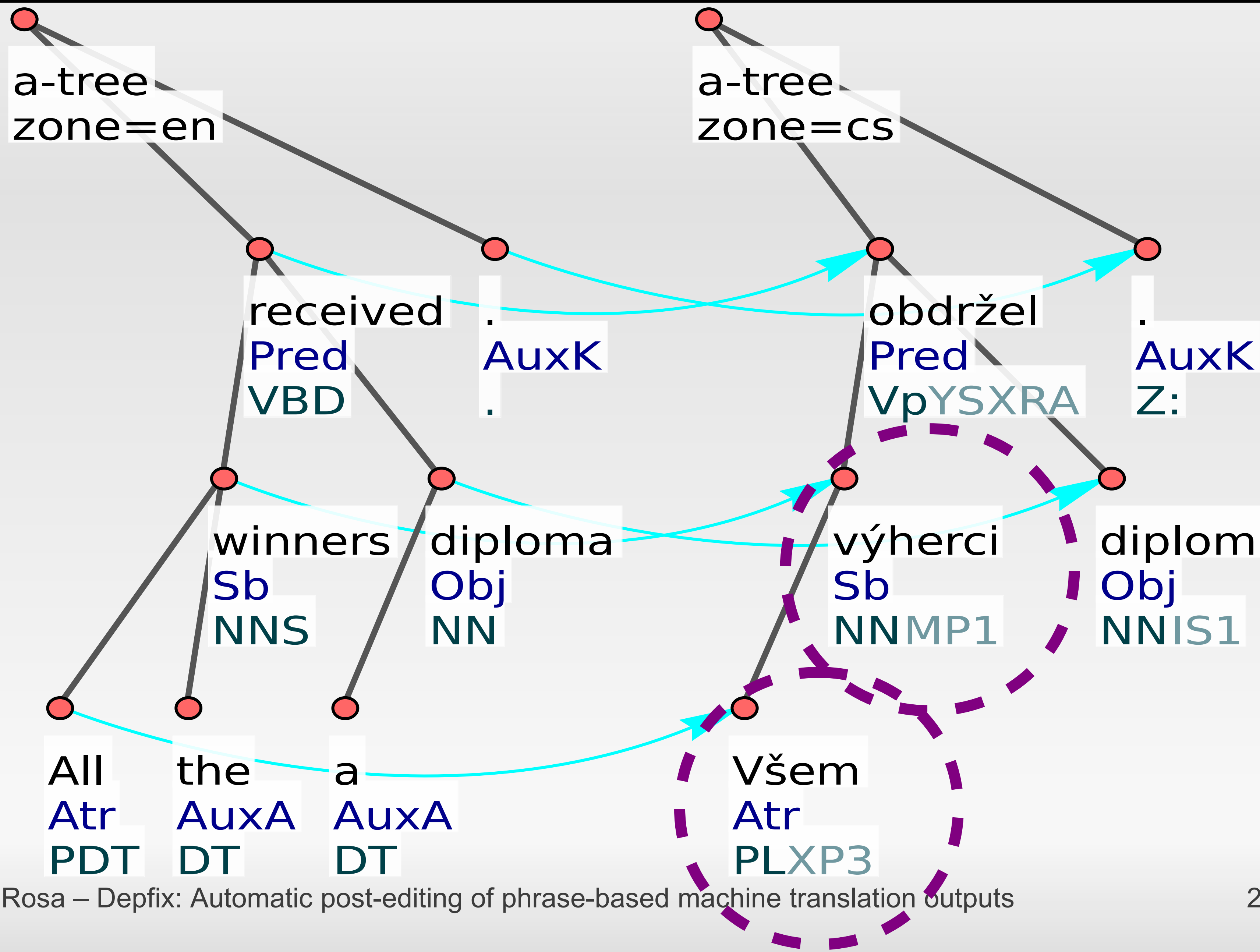
Subject → nominative



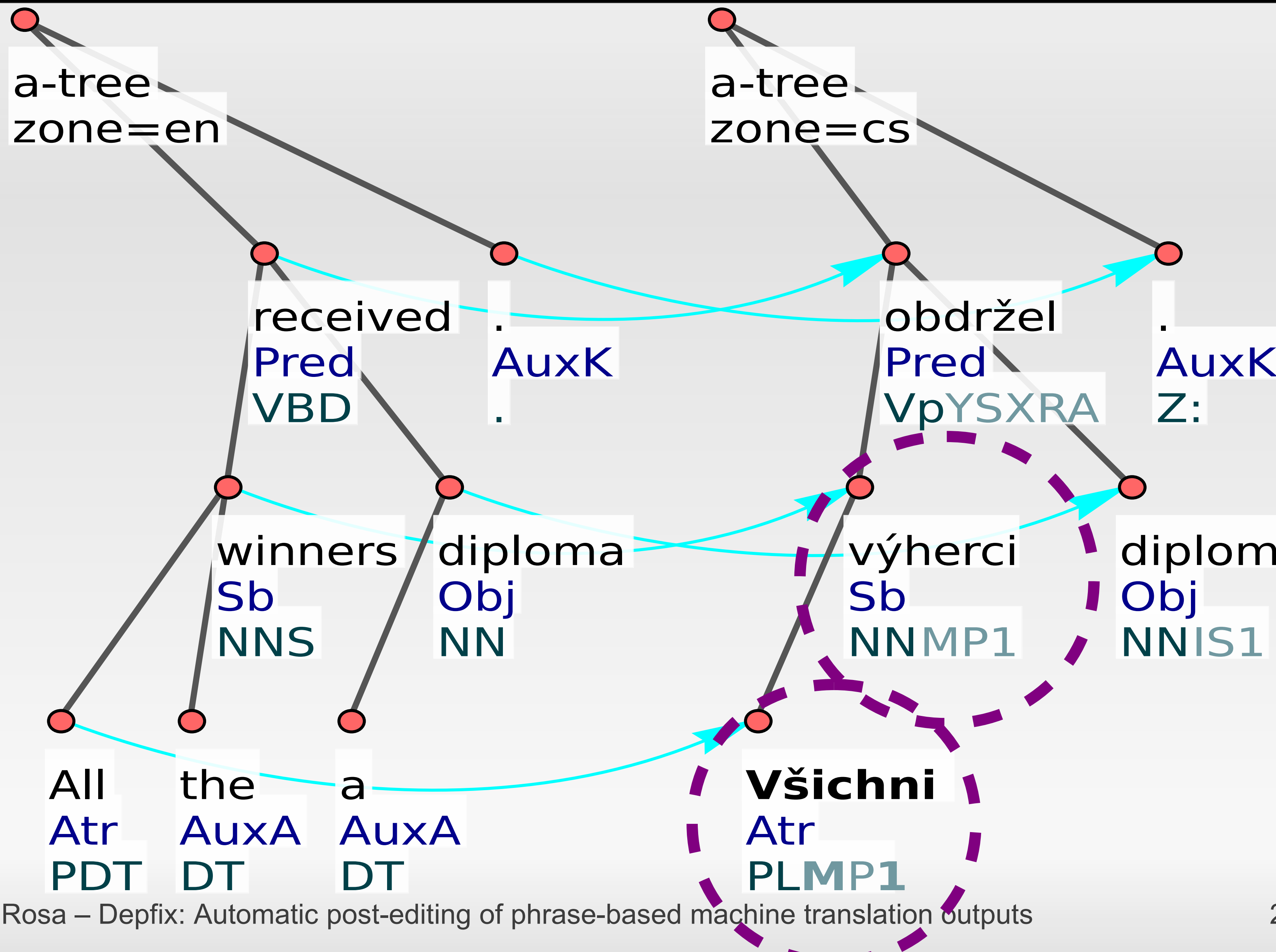
Všem výherci obdržel diplom.



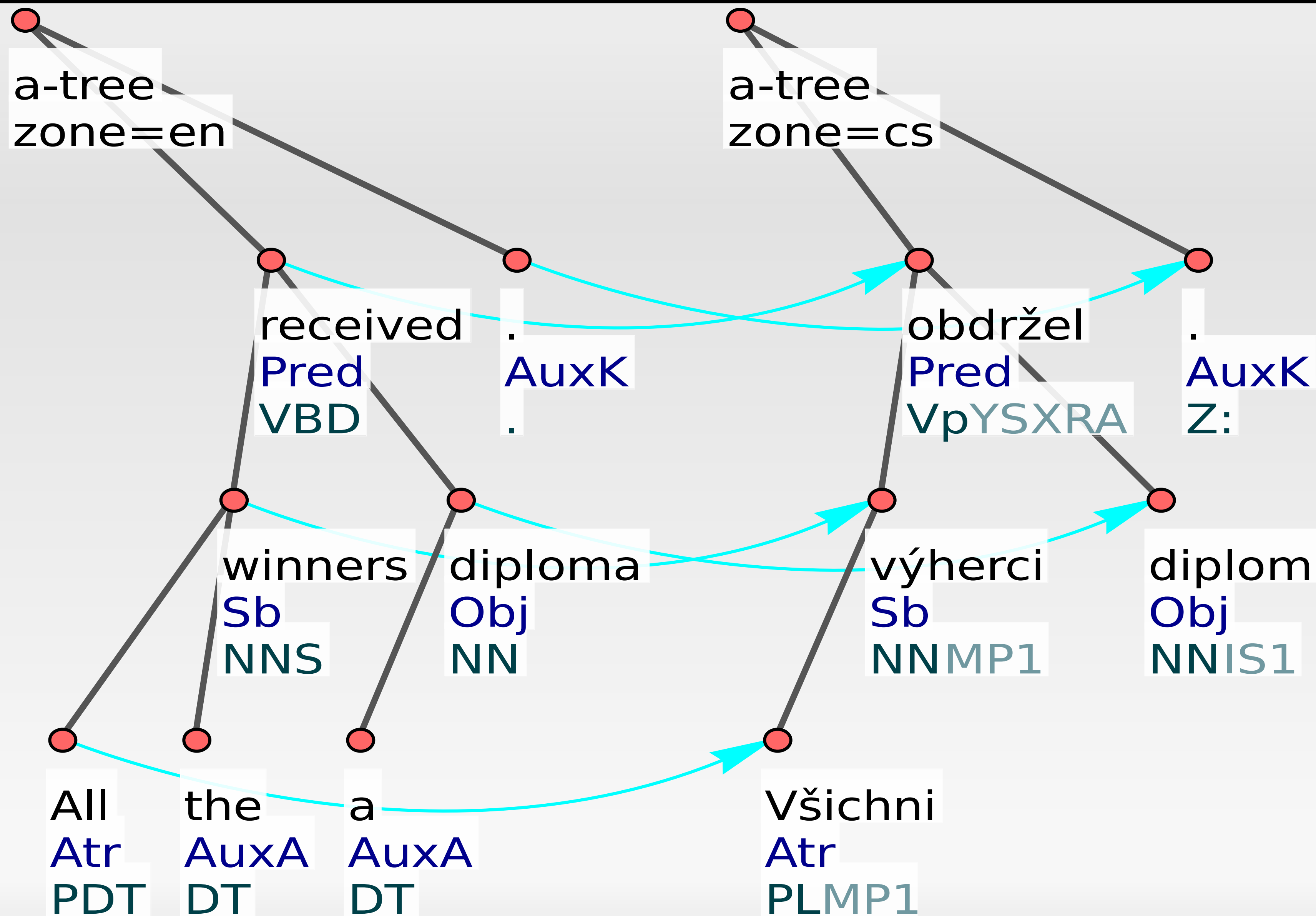
Noun-adjective agreement



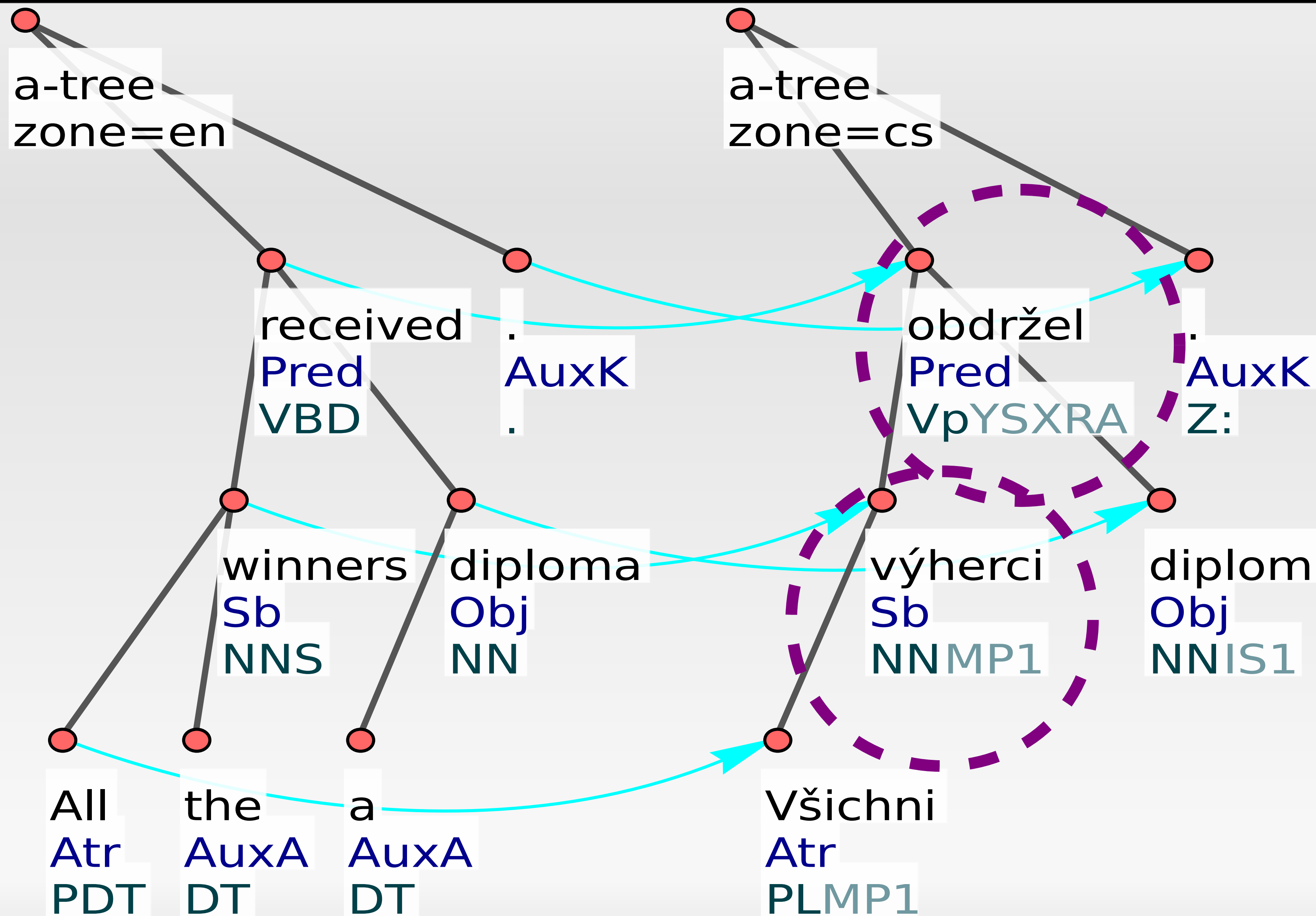
Agreement: gender, case (number)



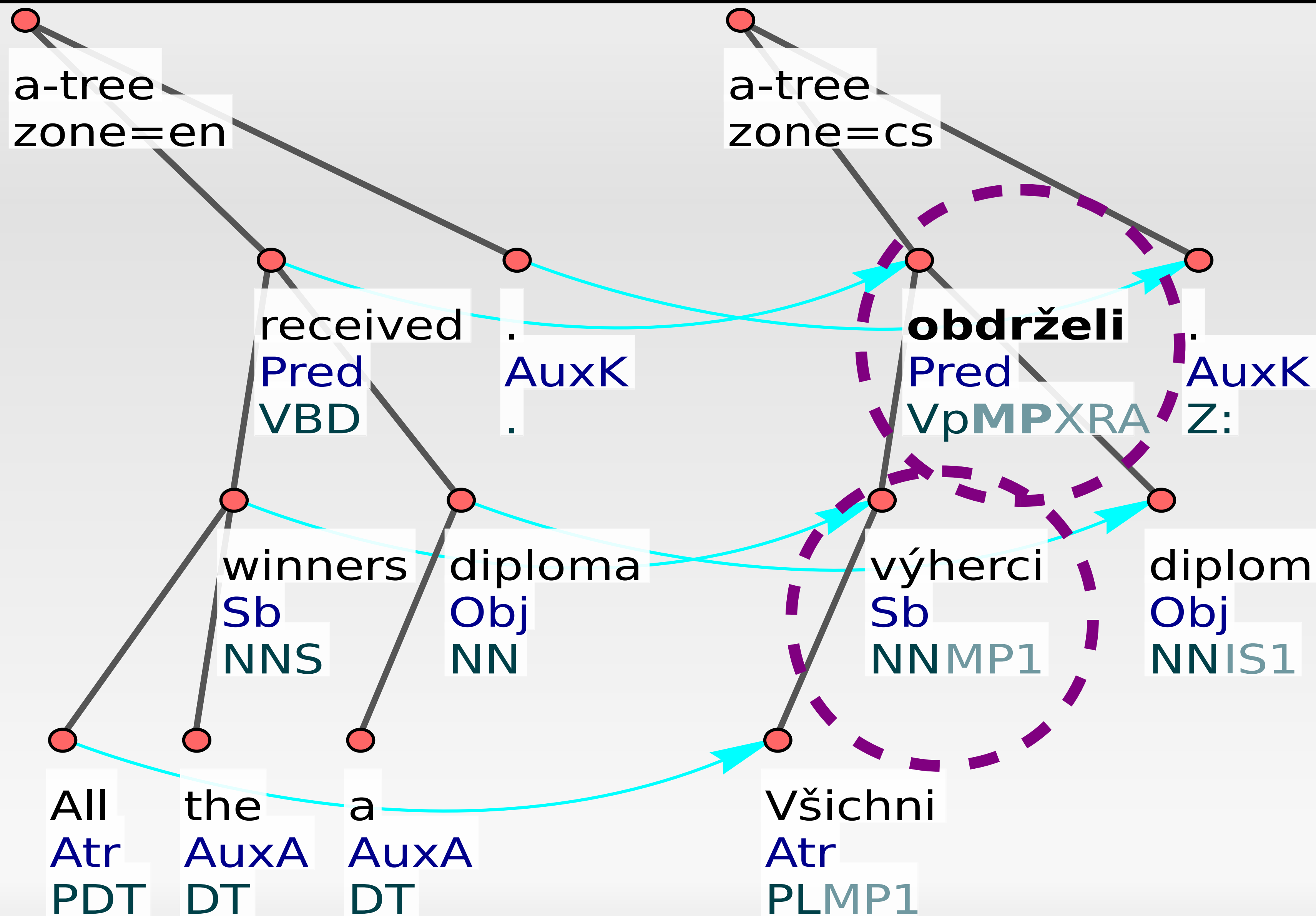
Všichni výherci obdržel diplom.



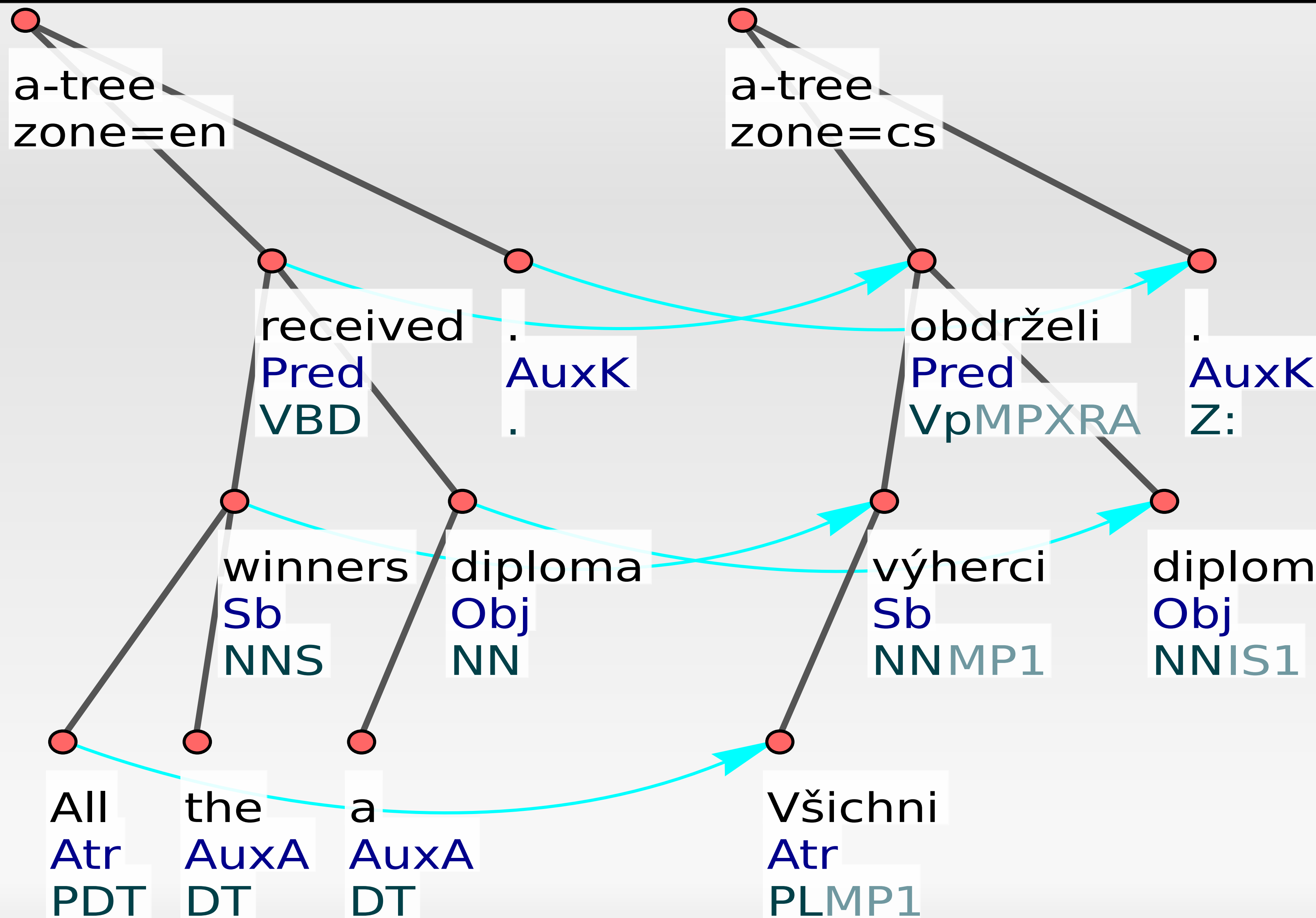
Subject-predicate agreement



Agreement: gender, num (person)



Všichni výherci obdrželi diplom.



Delving deeper...

- more corrections
  - an example of a cascade of corrections
  - **<u>translation of negation</u>**
  - correction of verb tense translation
- MSTperl parser
  - reimplementation of MST parser
  - adapted for parsing SMT outputs

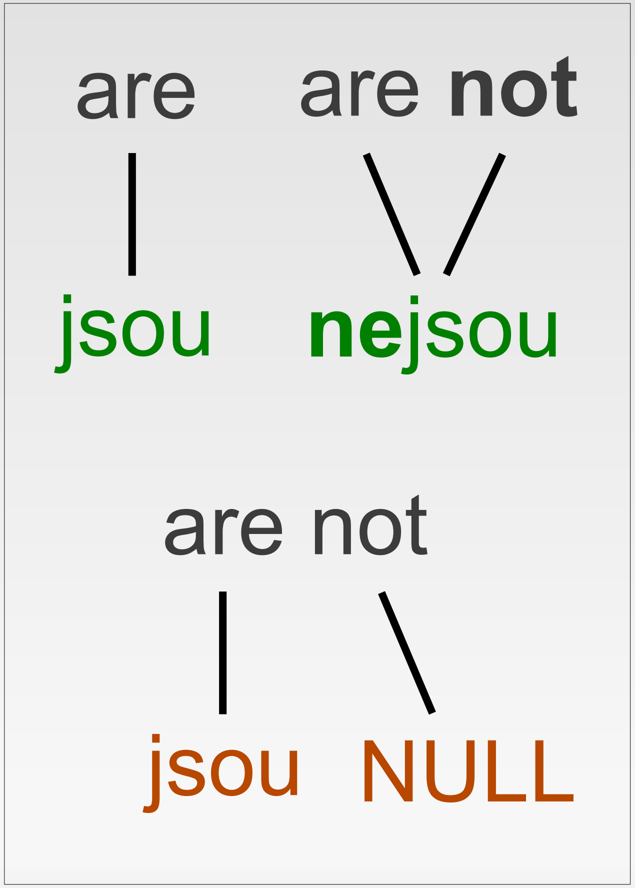
http://thedrivenclass.com/blog-detail/32876

Motivation

- *These **are not** actually errors.*
  - Moses: ***Jsou** to vlastně chyby.*
  - Gloss: *These **are** actually errors.*
  - Reference: ***Nejsou** to vlastně chyby.*
- *I **would not** cheat on you.*
  - Moses: *Já bych tě **podváděl**.*
  - Gloss: *I **would** cheat on you.*
  - Reference: *Já bych tě **nepodváděl**.*

Expressing negation

- default way in English: ***not* token**
  - *These are **not** actually errors.*
- default way in Czech: ***ne-* prefix**
  - ***Ne**jsou to vlastně chyby.*
- hard for word-alignment
- hard for PB SMT

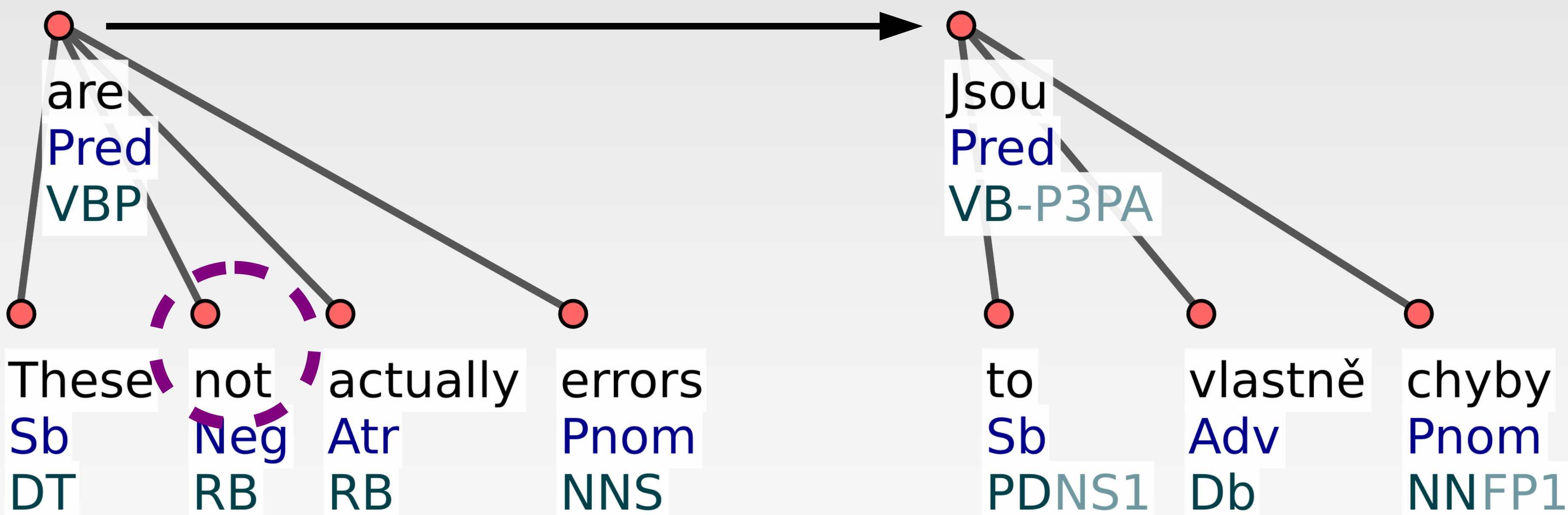


Actually, much more complex

- many ways to express negation (CS, EN)
  - negative particle (*not*), negative affix (*mis-*, *-less*), negative preposition (*without*)...
  - lexical means (*not happy ~ sad*)
  - differences between Czech and English
    - techniques based on word-alignment do badly
- the negation can be placed differently
  - usually it is the predicate heading the clause
    - but which part of it if it is multiword?
  - but not always

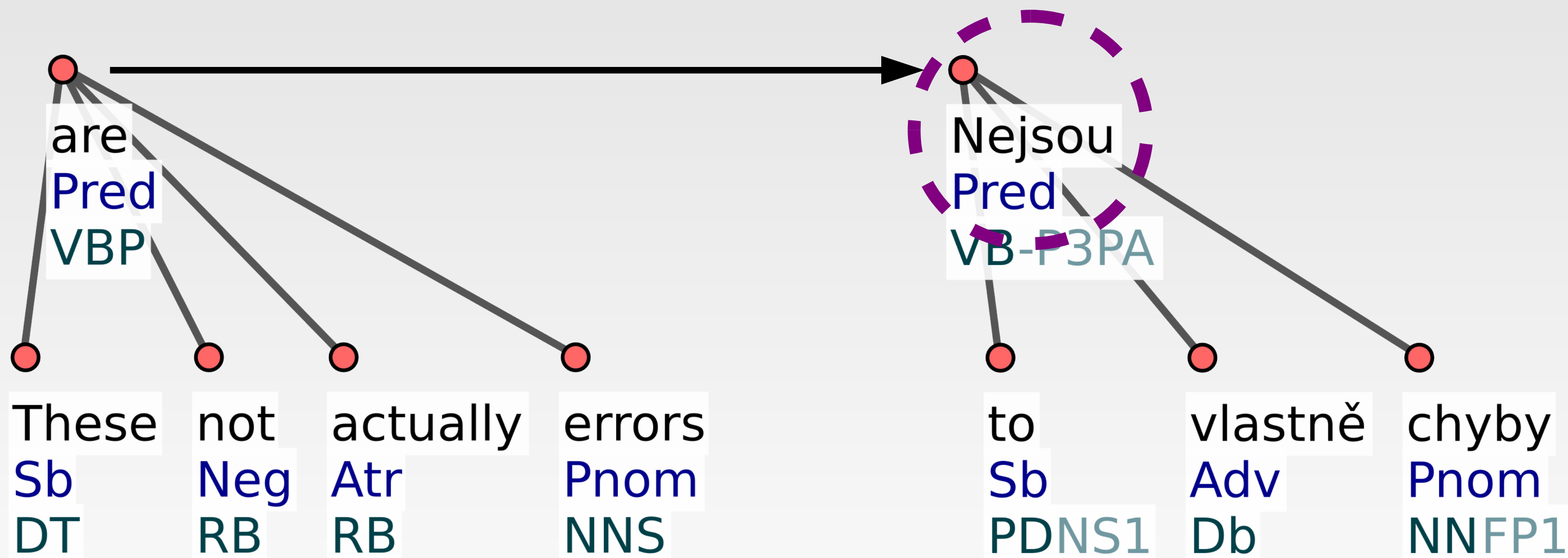
Detecting the problem

- English clause seems to be negative
- Czech clause seems to be positive



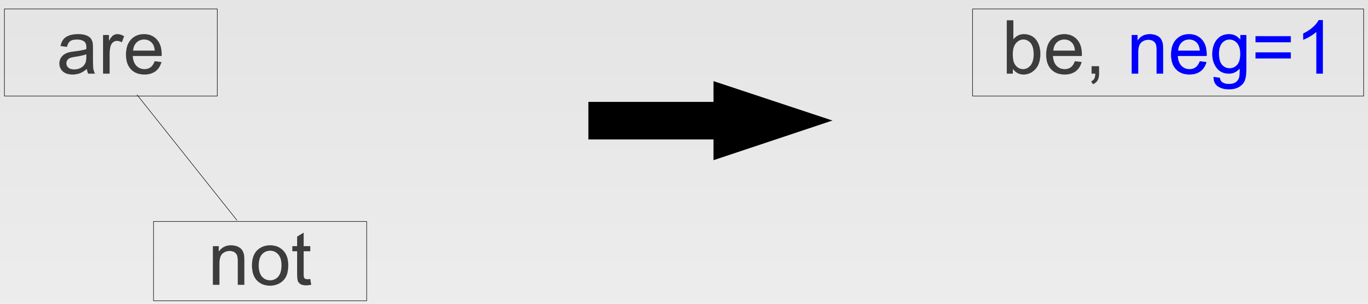
Fixing the problem

- find a place to put the negation (clause head)
- negate it (using tag & morpho generator)

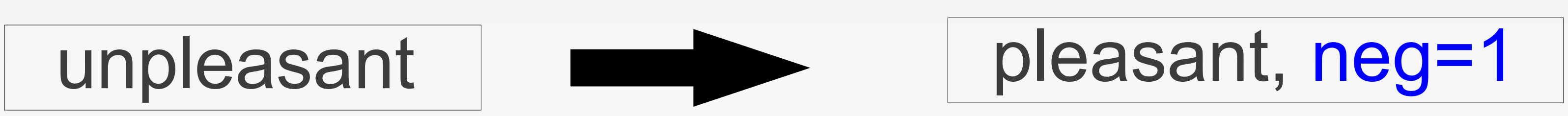


Deep syntactic analysis

- auxiliary nodes collapsed into values of attributes on parent nodes



- abstract from various ways of expressing negation (not, no, un-, in-,...)



Delving deeper...

- more corrections
  - an example of a cascade of corrections
  - correction of negation
  - **correction of verb tense translation**
- MSTperl parser
  - reimplementation of MST parser
  - adapted for parsing SMT outputs

Verb tense translation

- analyze the source (English) verb
- analyze the target (Czech) verb
- if they do not seem to match, change CS tense:
  - EN future * → CS future
  - EN past *; present perfect → CS past
  - EN present * → CS present
- avoid hard cases
  - conditionals, reported speech...

English verbs analysis

- parsing the verb form
  - all VB* and MD modifiers of the main verb
  - occasionally checking other modifiers (*to*)
- normalize to forms of *have*, *be* and *love*
- other words mark something
  - modality (modals, *have to, be (**un**)able to...*)
  - future (*will, going to* – careful: *was going to*)
  - conditionality (*would, should*)
  - past (*did*)

English verbs analysis

(present is the default – can be overidden by markers such as *did* or *will*)

- 'love' => [ ],
- 'loved' => ['past'],
- 'have loved' => ['perf'],
- 'be loving' => ['cont'],
- 'be loved' => ['pass'],
- 'had loved' => ['past', 'perf'],
- 'were loving' => ['past', 'cont'],
- 'were loved' => ['past', 'pass'],
- 'have been loving' => ['perf', 'cont'],
- 'have been loved' => ['perf', 'pass'],
- 'be being loved' => ['cont', 'pass'],
- 'had been loving' => ['past', 'perf', 'cont'],
- 'were being loved' => ['past', 'cont', 'pass'],
- 'had been loved' => ['past', 'perf', 'pass'],
- 'have been being loved' => ['perf', 'cont', 'pass'],
- 'had been being loved' => ['past', 'perf', 'cont', 'pass']

Delving deeper...

- more corrections
  - an example of a cascade of corrections
  - correction of negation
  - correction of verb tense translation
- **MSTperl parser**
  - reimplementation of MST parser
  - adaptation for parsing SMT outputs

MST parser

- Maximum Spanning Tree parser
- McDonald, Crammer, Pereira (2005)
  - Online large-margin training of dependency parsers
- McDonald, Pereira, Ribarov, Hajič (2006)
  - **Non-projective** dependency parsing using spanning tree algorithms
- discriminative, edge-local features
- MIRA learning algorithm

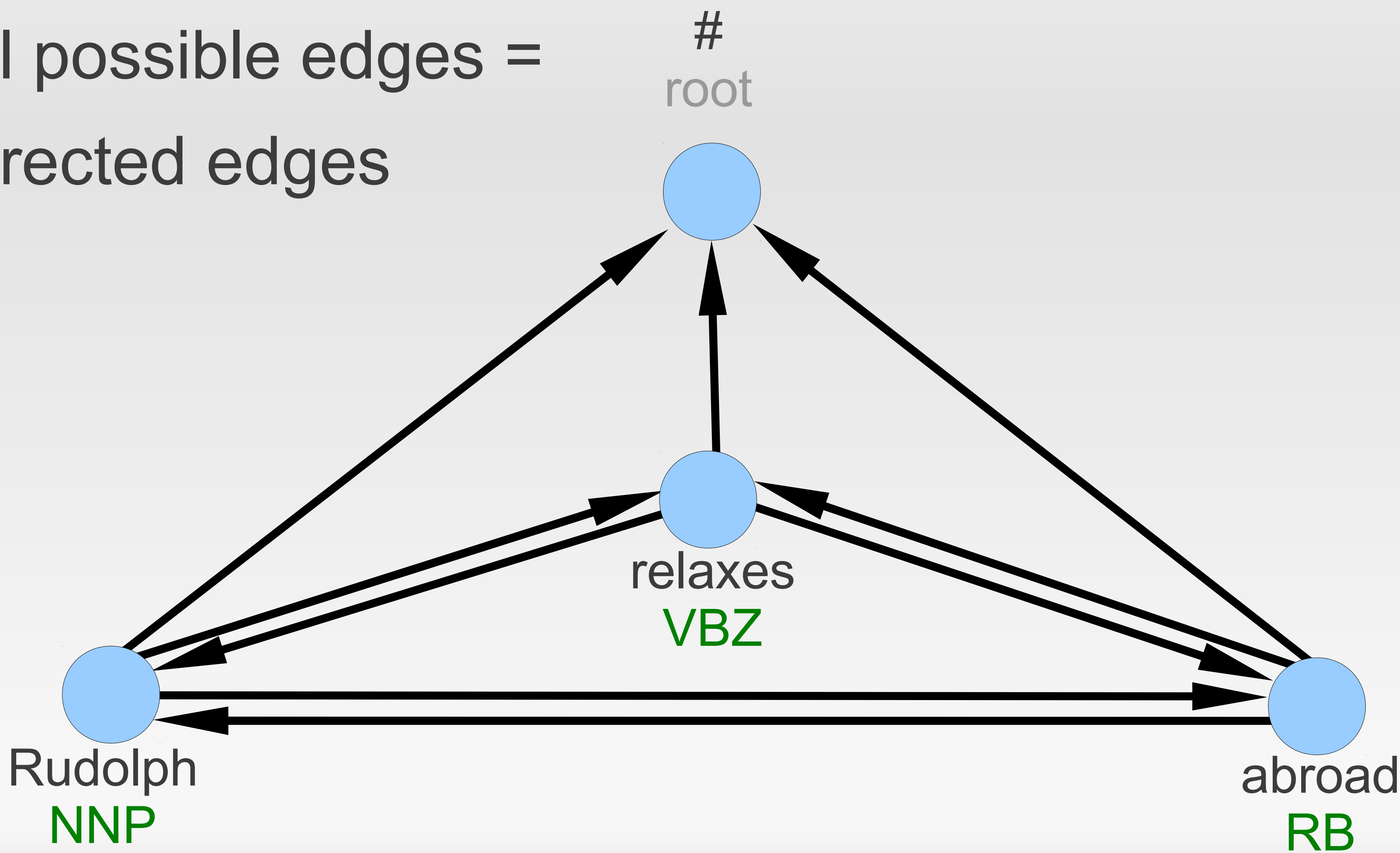
(1) Words and Tags

words = nodes

(2) (Nearly) Complete Graph

all possible edges =

directed edges

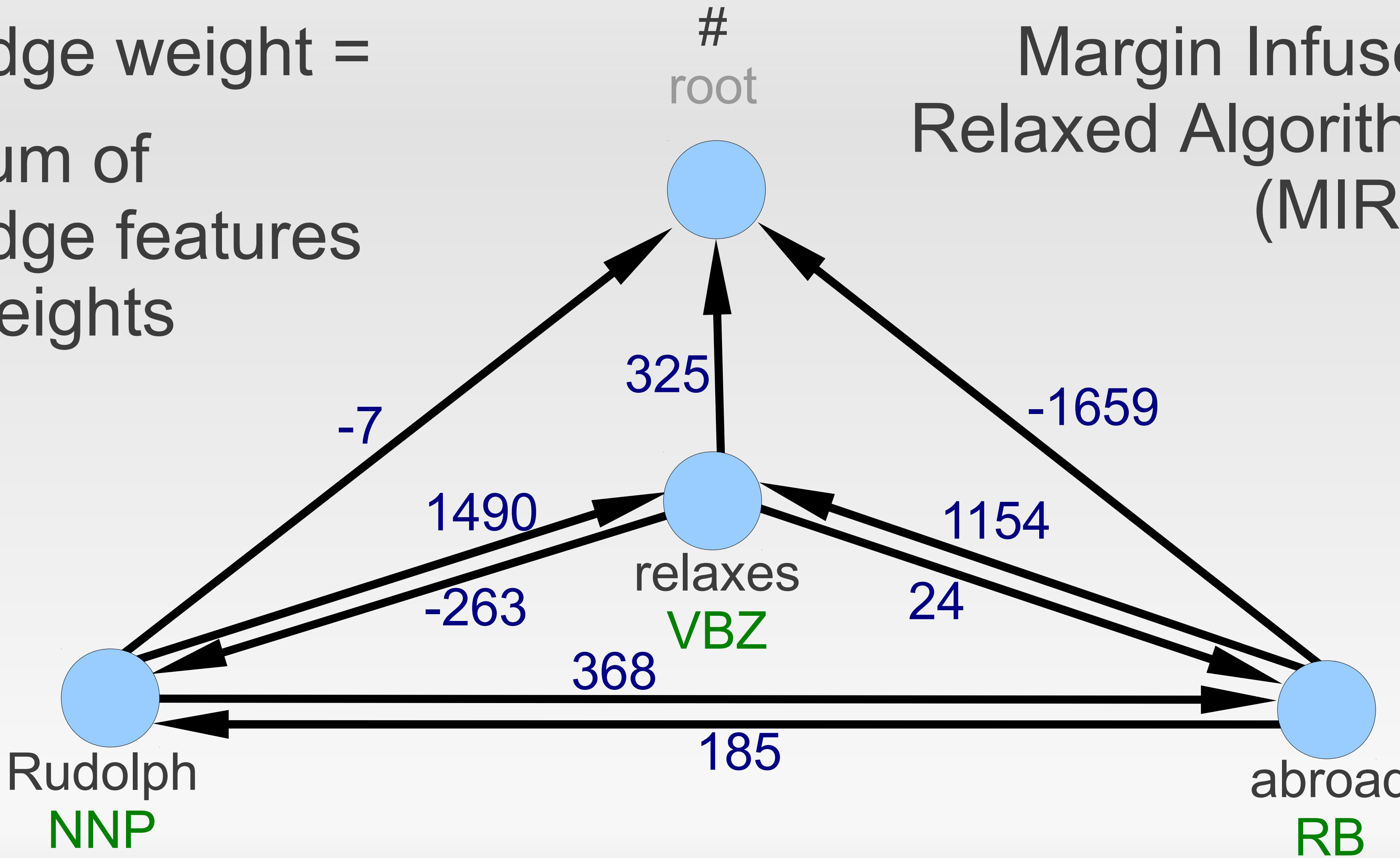


(3) Assign Edge Weights

edge weight =

sum of
edge features
weights

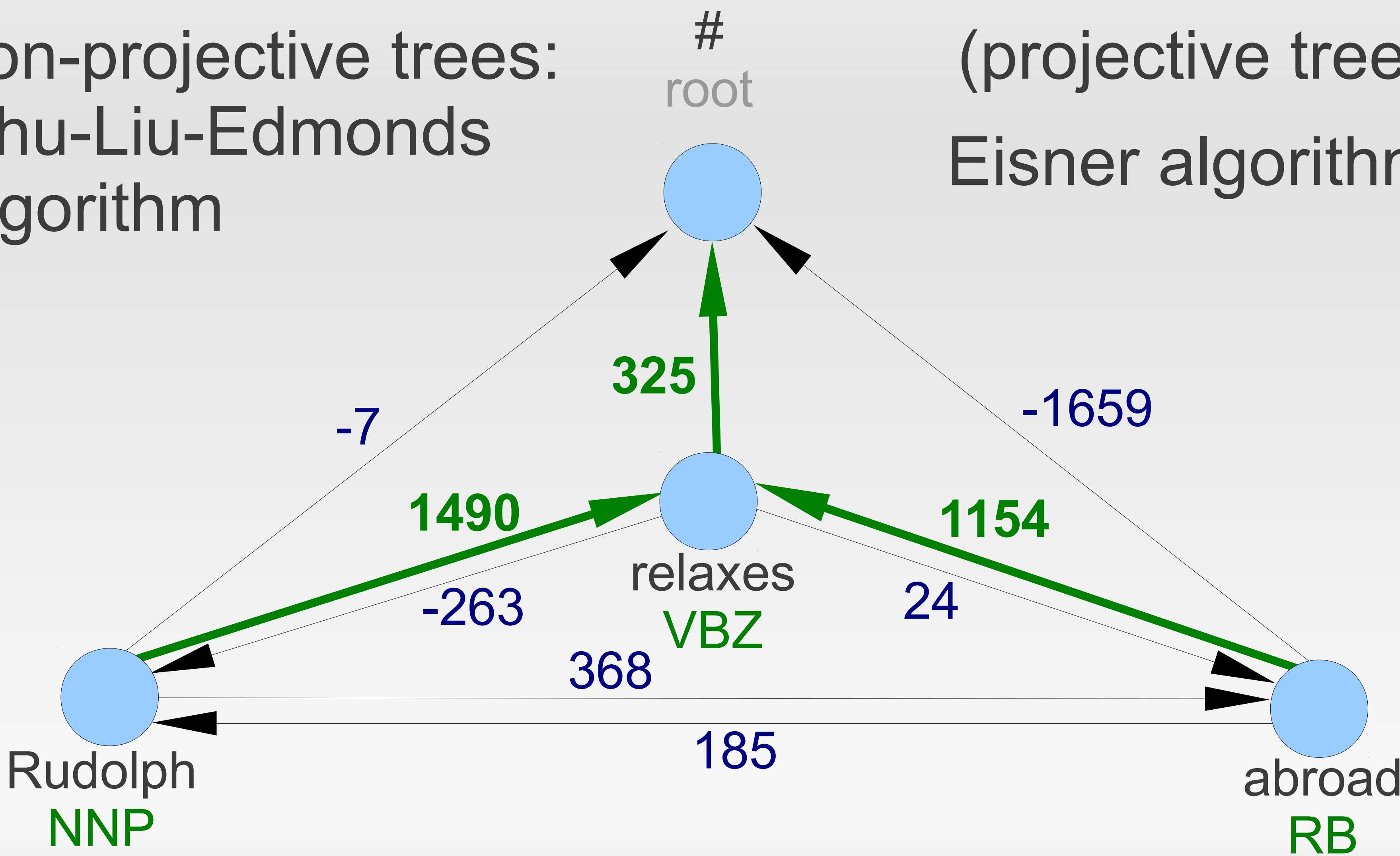
Margin Infused
Relaxed Algorithm
(MIRA)



(4) Maximum Spanning Tree

non-projective trees: Chu-Liu-Edmonds algorithm

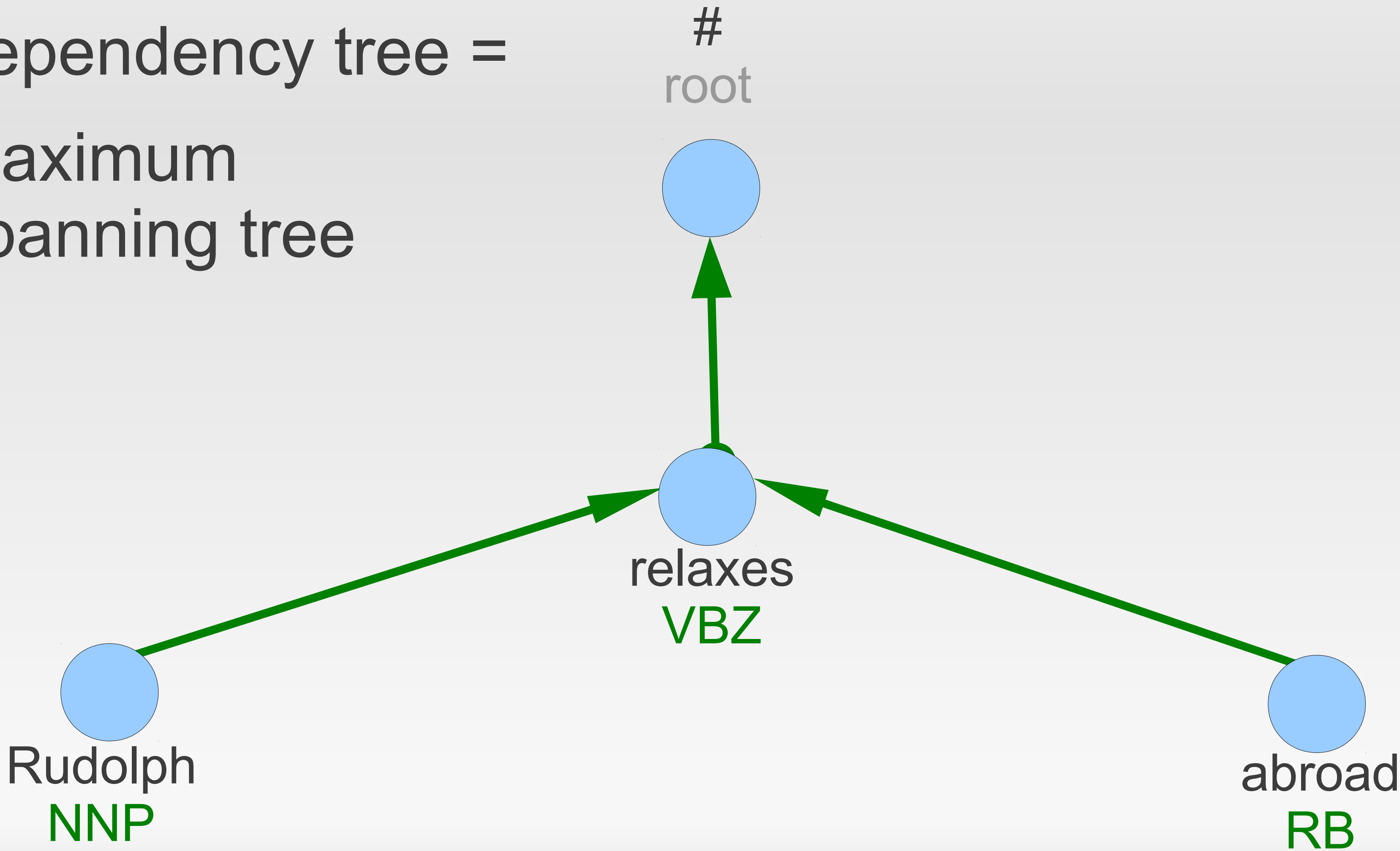
(projective trees: Eisner algorithm)



(5) Unlabeled Dependency Tree

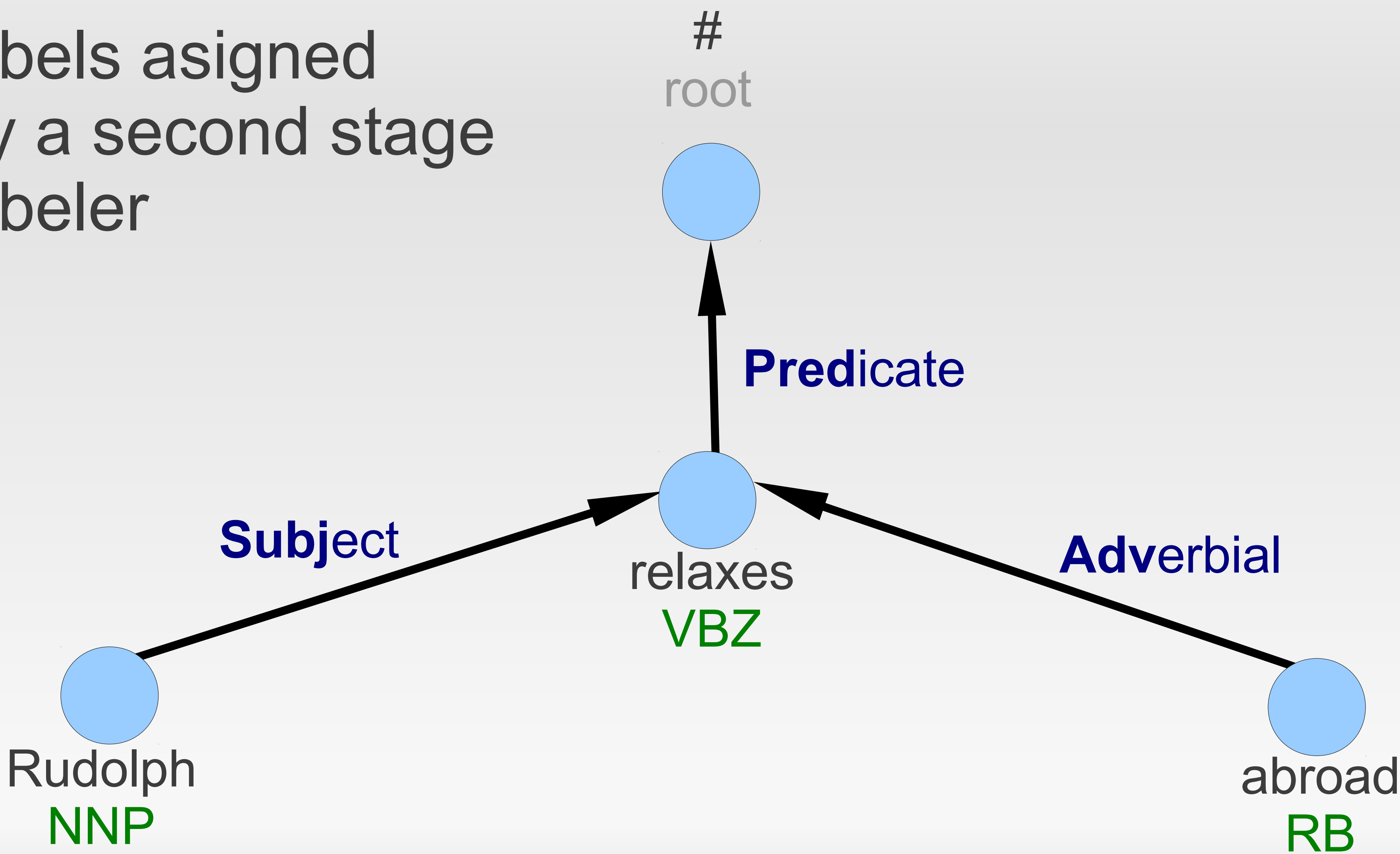
dependency tree =

maximum
spanning tree



(6) Labeled Dependency Tree

labels asigned
by a second stage
labeler



Parsing of SMT Outputs

- can be useful in many applications
  - automatic classification of translation errors
  - **automatic correction of translation errors (Depfix)**
  - 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?

MSTperl

- reimplementation of MST Parser in Perl
  - http://ufal.mff.cuni.cz/tools/mstperl-parser
  - first-order, non-projective
- adapted for SMT outputs parsing
  - worsening the training data
  - adding parallel information
  - manually boosting feature weights
  - exploiting large-scale data

MSTperl

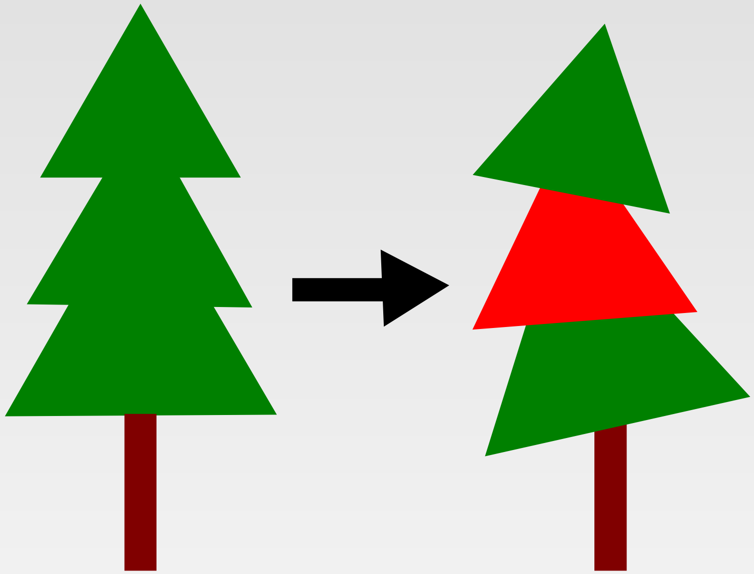
- reimplementation of MST Parser in Perl
  - http://ufal.mff.cuni.cz/tools/mstperl-parser
  - first-order, non-projective
- adapted for SMT outputs parsing
  - **worsening the training data**
  - adding parallel information
  - manually boosting feature weights
  - exploiting large-scale data

Parser Training Data

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

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
    - $\text{Rudolfovo}_{\text{NEUT}}$ $\text{auto}_{\text{NEUT}}$ je $\text{černé}_{\text{NEUT}}$.
  - SMT Czech
    - $\text{Rudolfova}_{\text{FEM}}$ $\text{auto}_{\text{NEUT}}$ je $\text{černý}_{\text{MASC}}$.

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

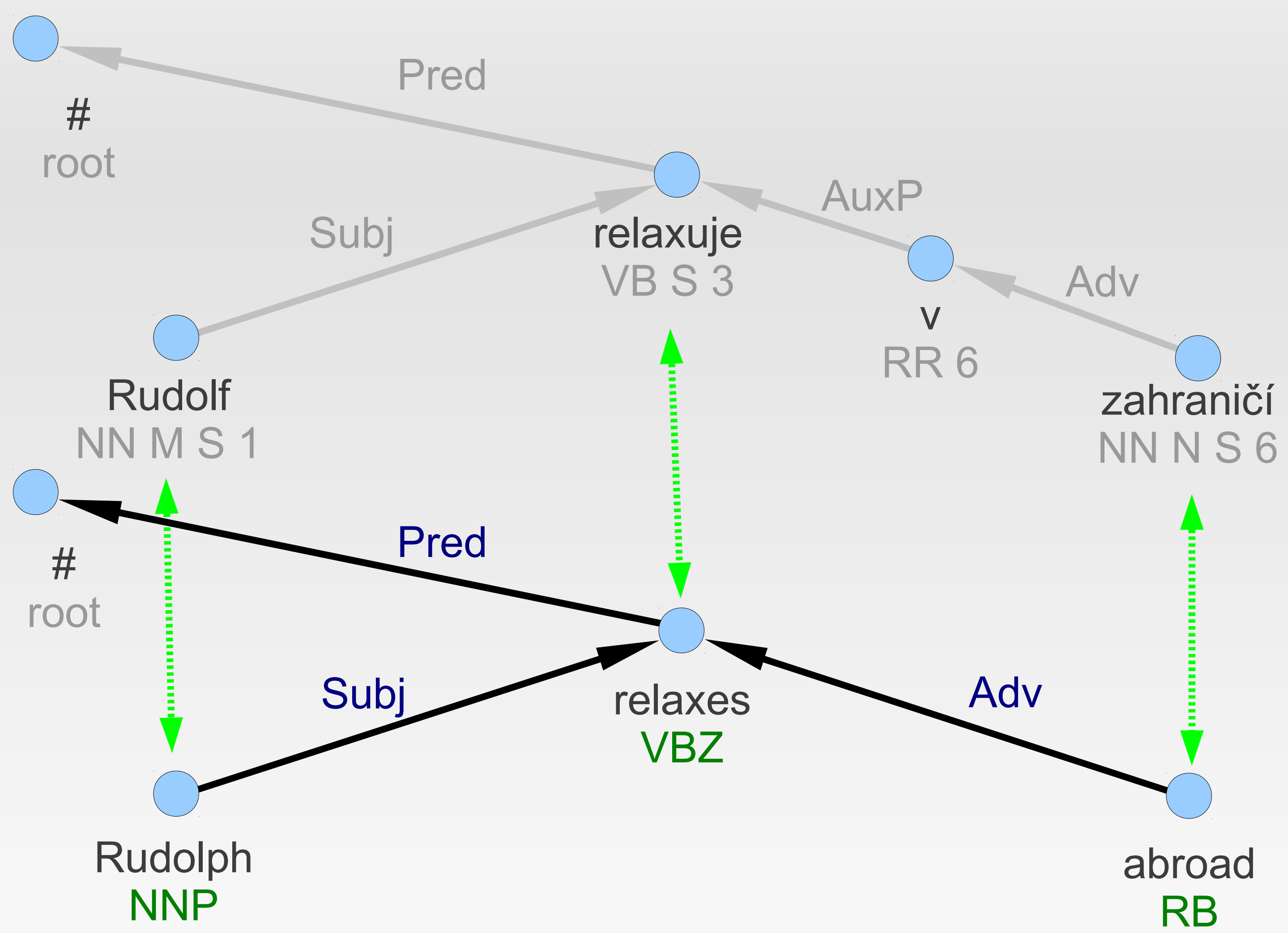
MSTperl

- reimplementation of MST Parser in Perl
  - http://ufal.mff.cuni.cz/tools/mstperl-parser
  - first-order, non-projective
- adapted for SMT outputs parsing
  - worsening the training data
  - **adding parallel information**
  - manually boosting feature weights
  - exploiting large-scale data

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



MSTperl

- reimplementation of MST Parser in Perl
  - http://ufal.mff.cuni.cz/tools/mstperl-parser
  - first-order, non-projective
- adapted for SMT outputs parsing
  - worsening the training data
  - adding parallel information
  - **manually boosting feature weights**
  - exploiting large-scale data

Manually boosting feature weights

- **aligned edge existence** is the key feature here
- observation: since the worsening is probably too mild, its weight is too low
  - edge exists: -0.57
  - edge does not exist: -0.83
  - missing aligned node(s): -0.67

Manually boosting feature weights

- **aligned edge existence** is the key feature here
- observation: since the worsening is probably too mild, its weight is too low
  - edge exists: **-0.57**
  - edge does not exist: -0.83
  - missing aligned node(s): -0.67
- experiment: manually increase its weight
  - edge exists: **-0.25**

Manually boosting feature weights

- **aligned edge existence** is the key feature here
- observation: since the worsening is probably too mild, its weight is too low
  - edge exists: **-0.57**
  - edge does not exist: -0.83
  - missing aligned node(s): -0.67
- experiment: manually increase its weight
  - edge exists: **-0.25**
- success – manual changing of weights feasible

MSTperl

- reimplementation of MST Parser in Perl
  - http://ufal.mff.cuni.cz/tools/mstperl-parser
  - first-order, non-projective
- adapted for SMT outputs parsing
  - worsening the training data
  - adding parallel information
  - manually boosting feature weights
  - **exploiting large-scale data**

Exploiting large-scale data

- exploiting large-scale parsed data (CzEng) to provide additional lexical features
- lexical features are important for the parser
- CzEng has 10 times more word types (lemmas) than PCEDT (400k vs. 40k)
- training the parser on whole CzEng infeasible
- new feature: pointwise mutual information

$$PMI'(parent, child) = \log \frac{count([parent, child])}{count([parent, *]) \cdot count([*, child])}$$

Direct Evaluation: by Inspection

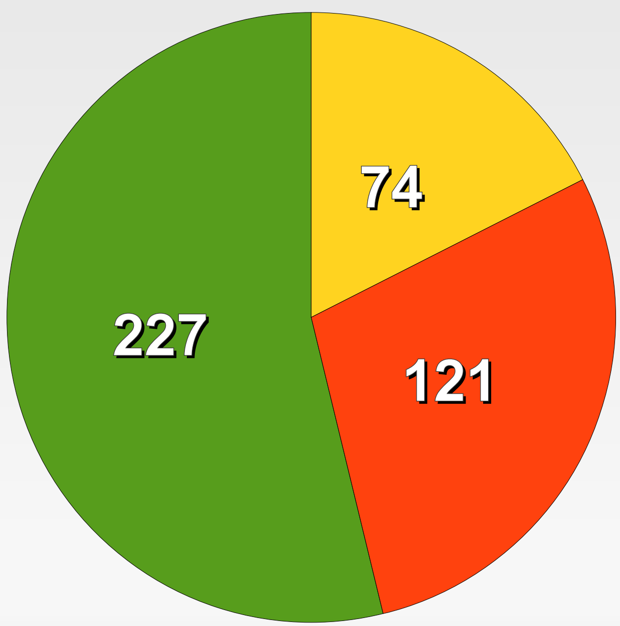
- manual inspection of several parse trees
  - comparing baseline and adapted parser ouputs
- examples of improvements:
  - subject identification even if not in nominative case
  - adjective-noun dependence identification even if agreement violated (gender, number, case)
- hard to do reliably
  - trying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

Indirect Evaluation: in Depfix

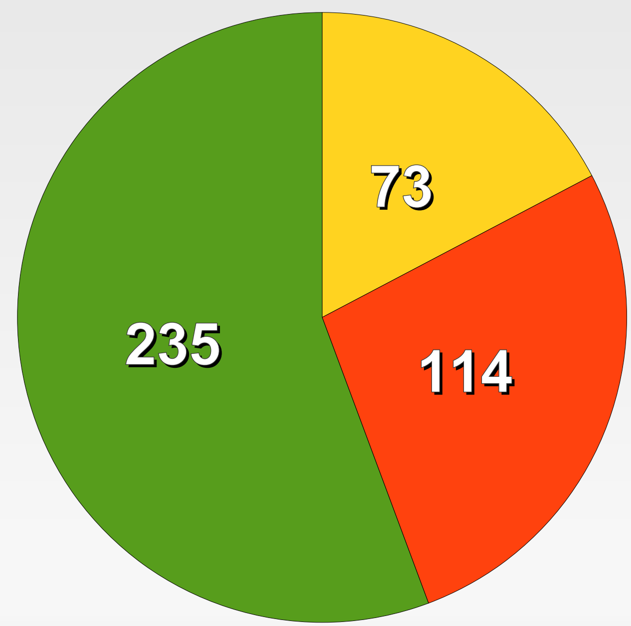
- run Depfix with
  - baseline 1: the original McDonald's MST parser
  - baseline 2: basic MSTperl (without the adaptations)
  - adapted MSTperl
- manual evaluation of adapted MSTperl versus the two baseline parsers
  - how many sentences come out better from Depfix using adapted MSTperl than from Depfix using a baseline parser

Indirect Evaluation: in Depfix

- improvements and deteriorations in Depfix:
  - adapted MSTperl vs original McDonald's MST Parser i'm parsin' it



  - adapted MSTperl vs basic MSTperl



Conclusion

- automatic post-editing of SMT is possible
  - “easy” with using linguistic analysis and generation
  - adapting the parser for SMT outputs also helps
- rule-based system for English→Czech
  - achieves improvements across SMT systems
- machine-learned system (**now** English→Czech)
  - could learn more fine-grained rules
  - could be easily extended to other languages
    (if we have analysis and generation)

Thank you for your attention

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Depfix:

Automatic post-editing of phrase-based machine translation outputs

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

For this presentation and other information, please visit:

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