

Rudolf Rosa

diplomová práce

Automatická post-editace výstupů frázového strojového překladu (Depfix)

Automatic post-editing
of phrase-based machine translation outputs

Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*

Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?



Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?
 - *Všem výhercům obdržel diplom.*



Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?
 - *Všem výhercům obdržel diplom.*
- Co by se nám líbilo více?



Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?
 - *Všem výhercům obdržel diplom.*
- Co by se nám líbilo více?
 - *Všichni výherci obdrželi diplom.*



Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?
 - *Všem výhercům obdržel diplom.*
- Co na to Depfix?



Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?
 - *Všem výhercům obdržel diplom.*
- Co na to Depfix?
 - vezme zdrojovou větu a výstup Mosese



Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?
 - **Všem výhercům obdržel** diplom.
- Co na to Depfix?
 - vezme zdrojovou větu a výstup Mosese
 - najde chyby



Motivační příklad

- Zdroj (WMT 2010):
 - *All the winners received a diploma.*
- Co nám dá Moses?
 - *Všem výhercům obdržel diplom.*
- Co na to Depfix?
 - vezme zdrojovou větu a výstup Mosese
 - najde chyby, opraví je, a vydá správný překlad
 - *Všichni výherci obdrželi diplom.*



Obsah

- Motivační příklad
- Depfix krok za krokem
- Úpravy použitých nástrojů, zejména parseru
- Vyhodnocení
- Shrnutí

Depfix krok za krokem

- Lingvistická analýza vstupu
- Pravidlové a statistické opravy chyb
- Roviny dle Pražského závislostního korpusu
 - M-rovina
 - A-rovina
 - T-rovina
- Implementováno ve frameworku Treex
 - přijata Treexová implementace rovin

M-rovina

- Analýza: lemmata, tagy, word-alignment
- Opravy: vokalizace předložek, kapitalizace

M-rovina

- Analýza: lemmata, tagy, word-alignment
- Opravy: vokalizace předložek, **kapitalizace**
 - *Zdroj*: director of the best hotel in **Pec**, Karel **Rada**
 - *Moses*: ředitel nejlepší hotel v **peci**, Karel **rada**
 - *Depfix*: ředitel nejlepší hotel v **Peci**, Karel **Rada**

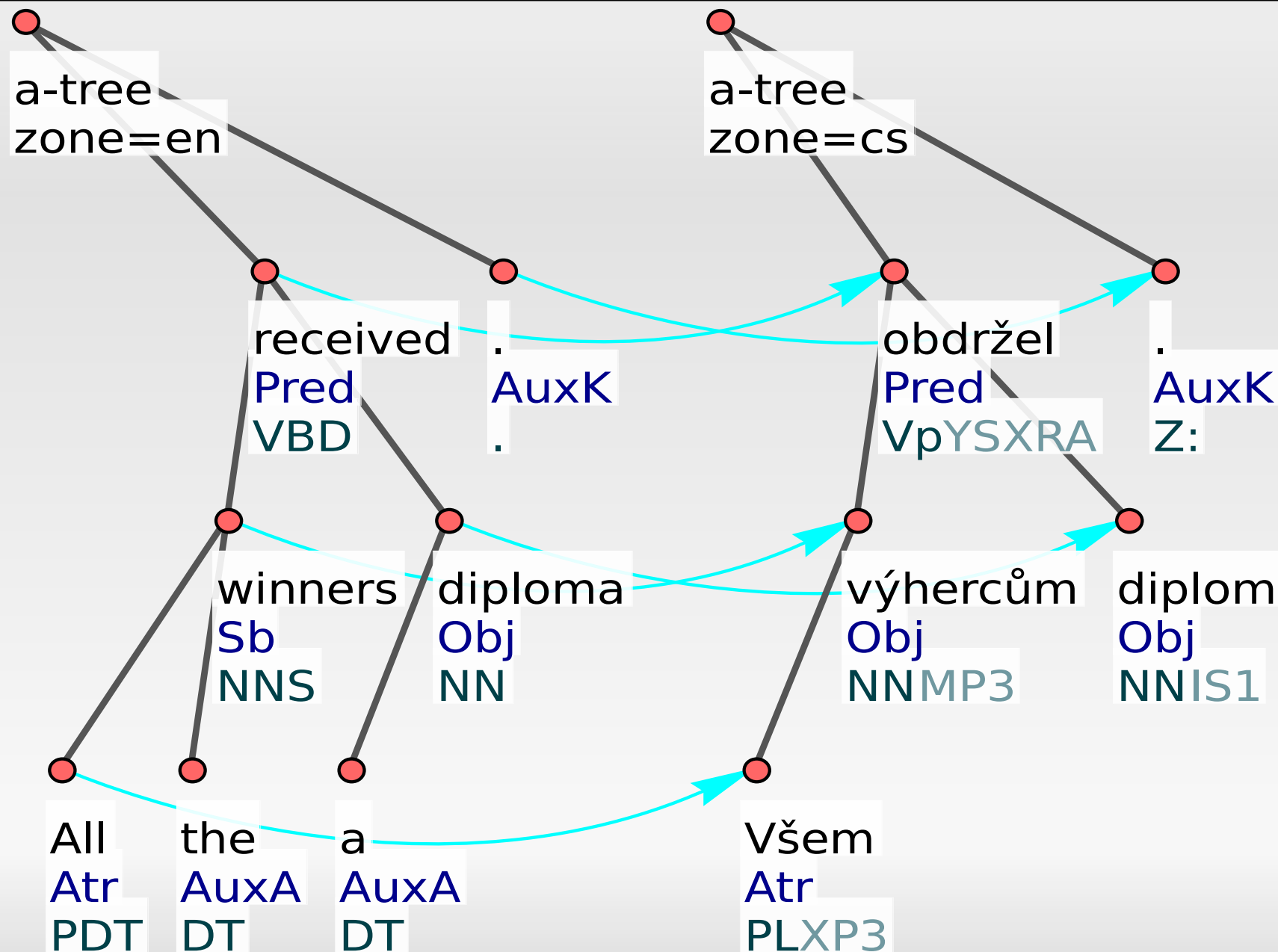
M-rovina

- Analýza: lemmata, tagy, word-alignment
- Opravy: vokalizace předložek, **kapitalizace**
 - *Zdroj*: director of the best hotel in **Pec**, Karel **Rada**
 - *Moses*: ředitel **nejlepší hotel** v **peci**, Karel **rada**
 - *Depfix (M)*: ředitel nejlepší hotel v **Peci**, Karel **Rada**
 - *Depfix*: ředitel **nejlepšího hotelu** v **Peci**, Karel **Rada**

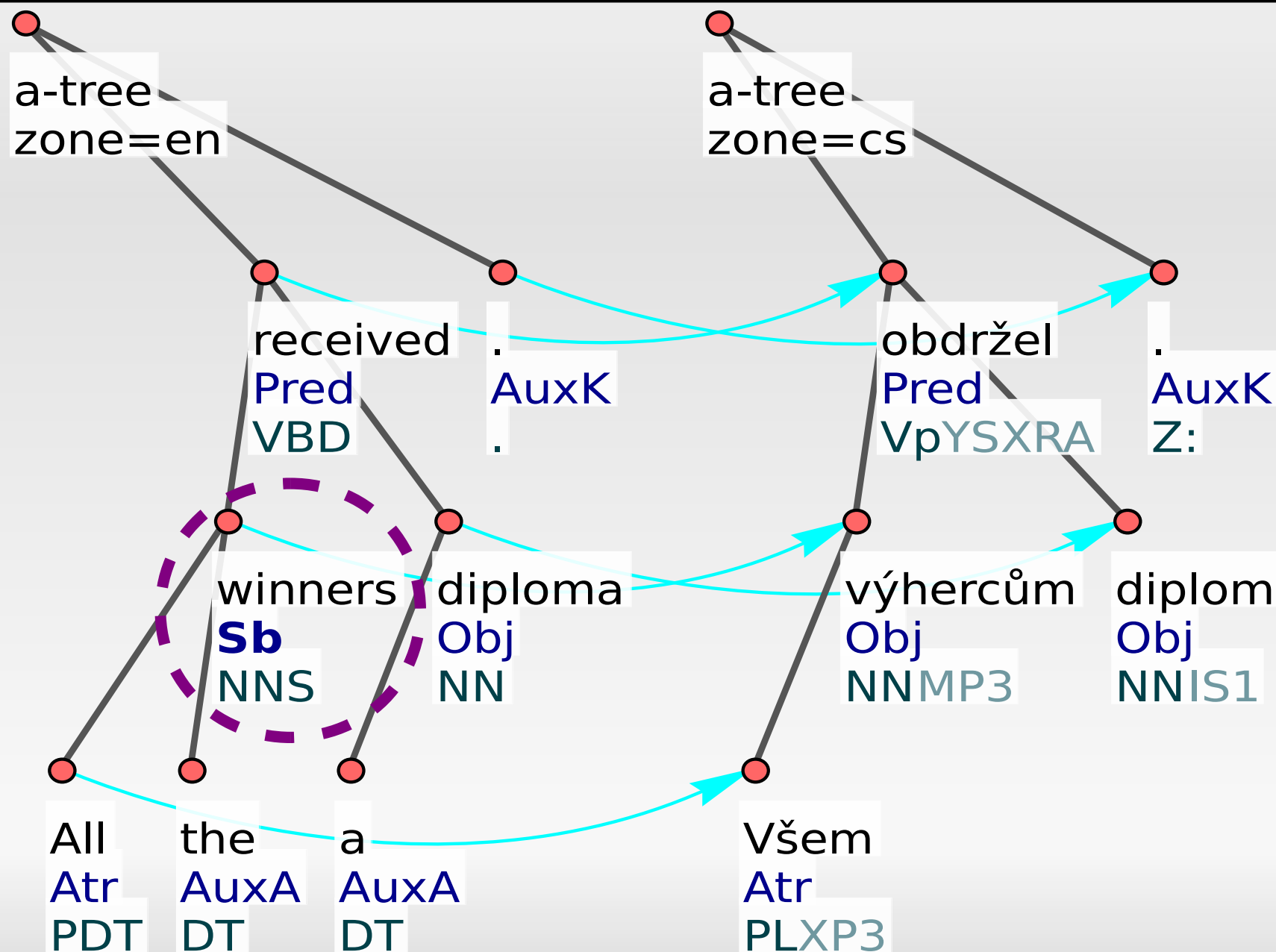
A-rovina

- Analýza: závislostní stromy, analytické funkce
- Opravy:
 - **morfologické shody:** předložka se substantivem, podmět s přísudkem, substantivum s adjektivem
 - **transfer významu do morfologie:** podmět, nevyjádřený podmět, přivlastňování, pasivum

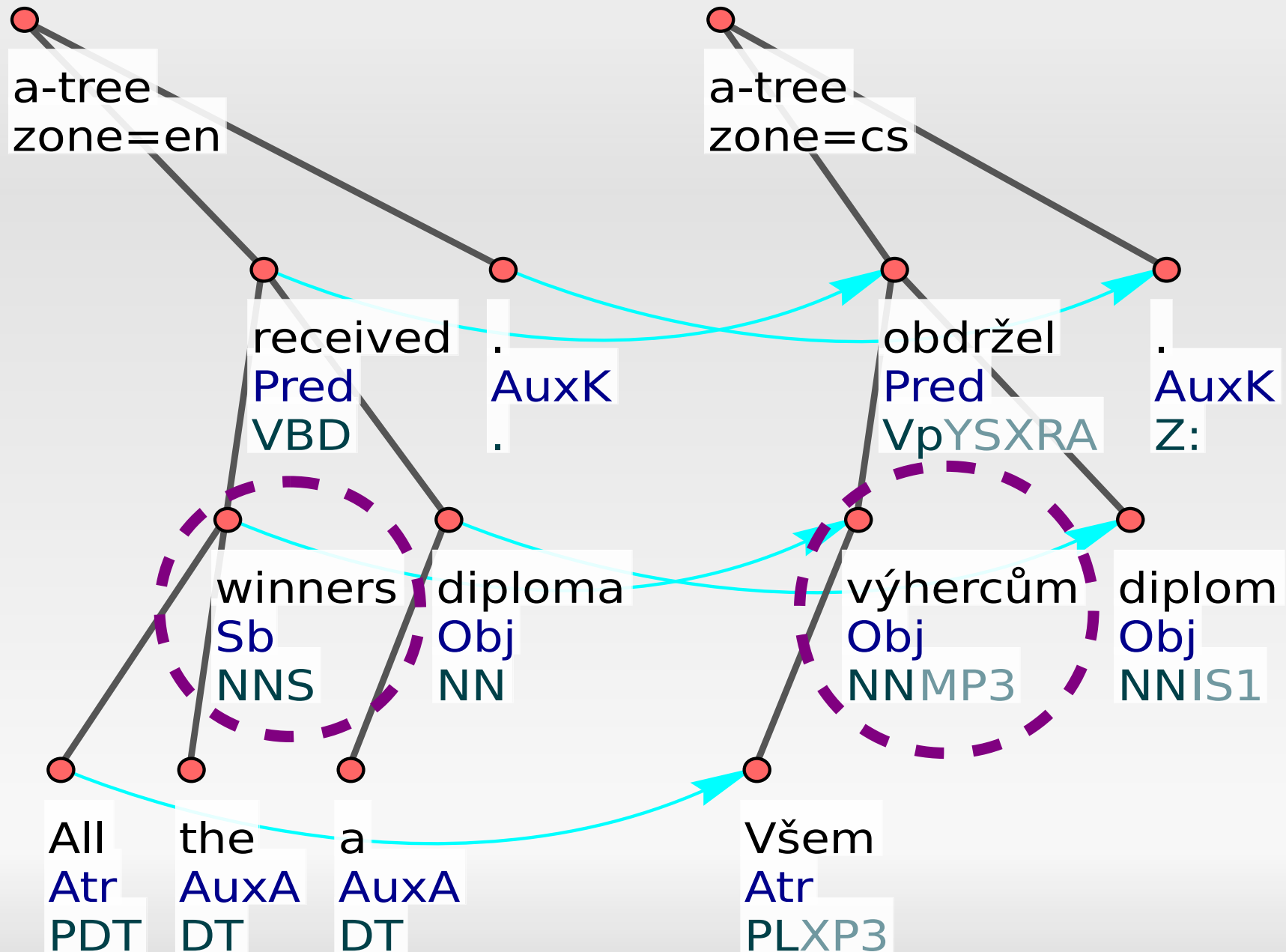
Všem výhercům obdržel diplom.



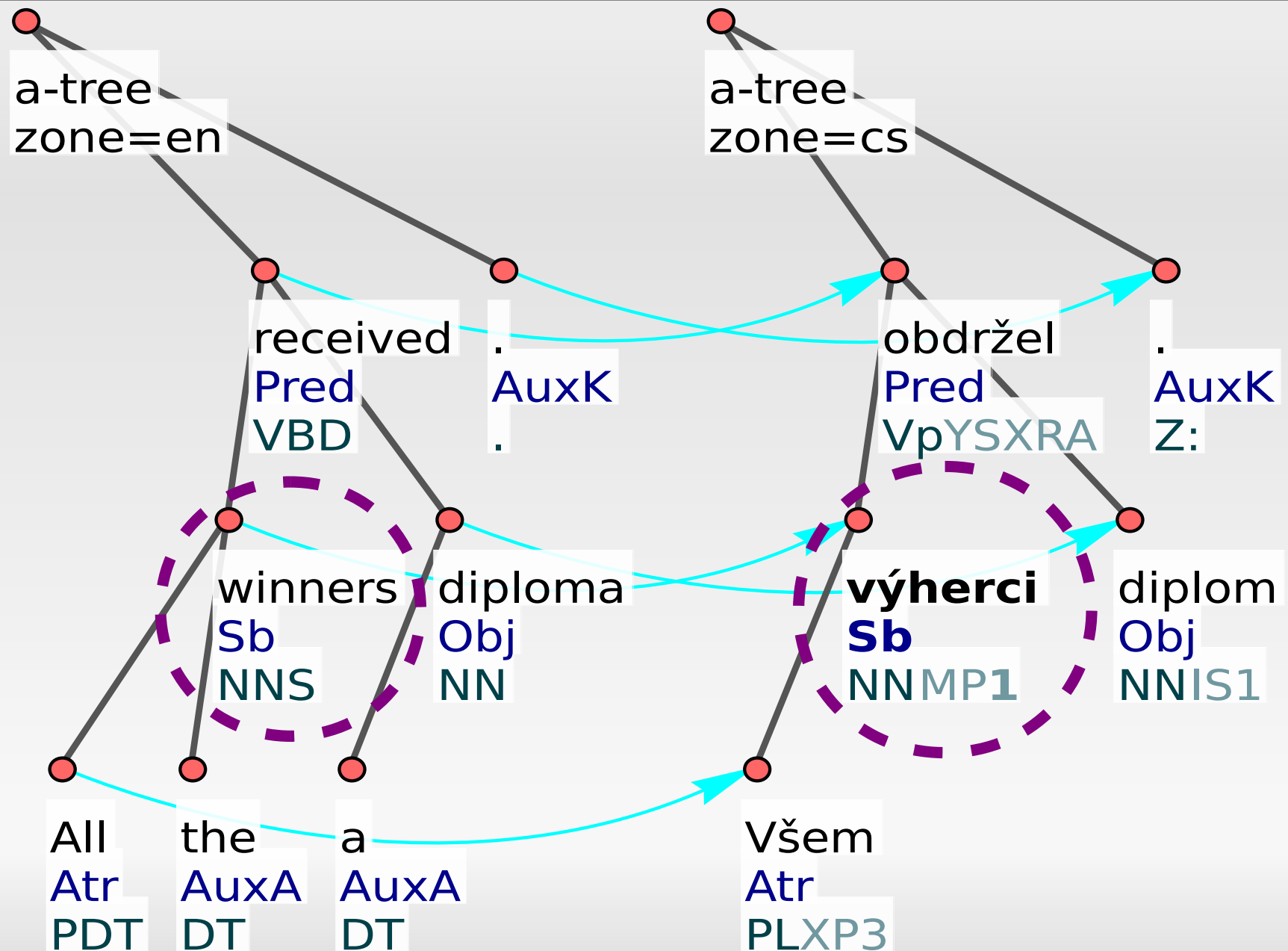
Transfer významu: podmět



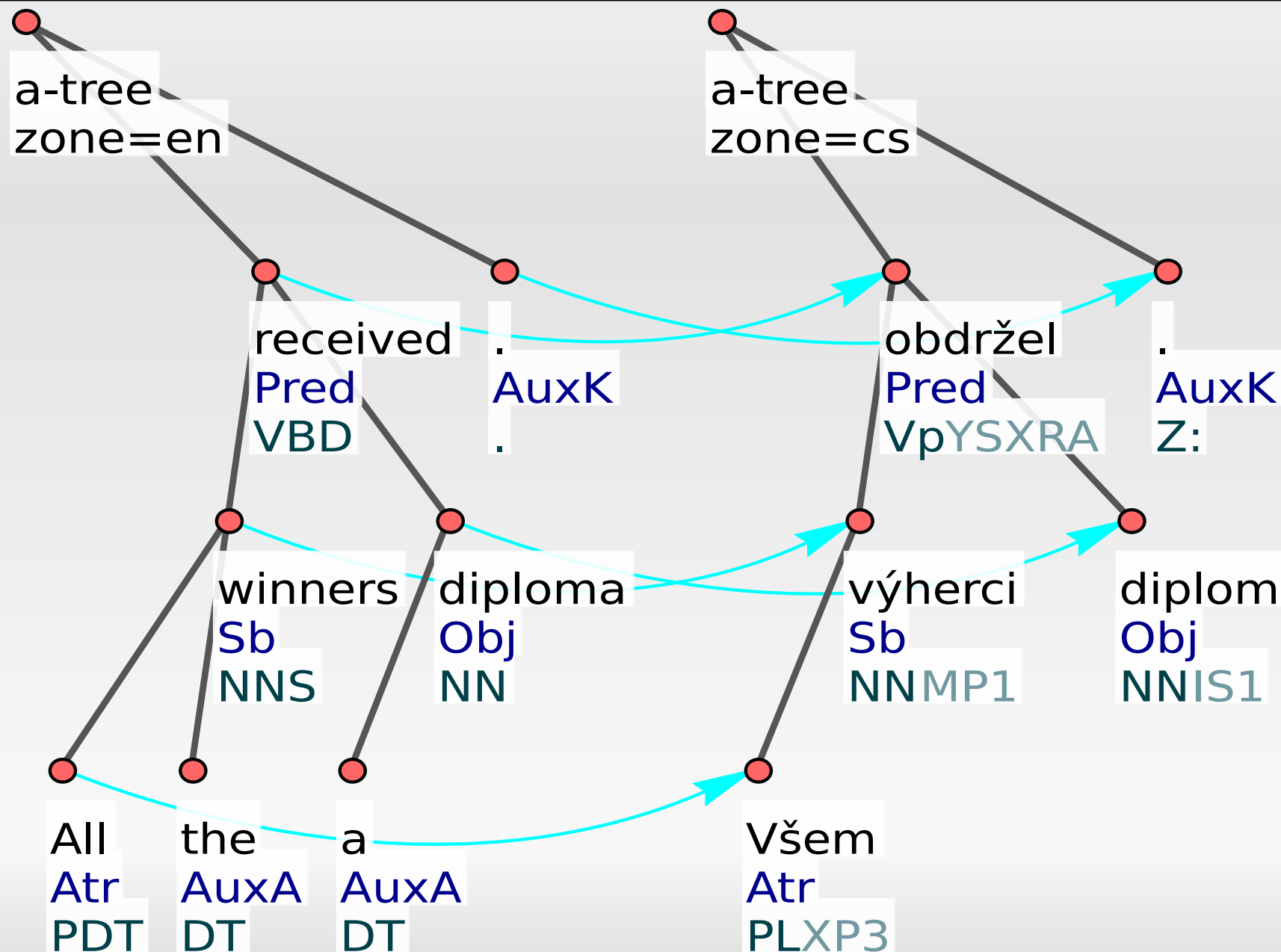
Transfer významu: podmět



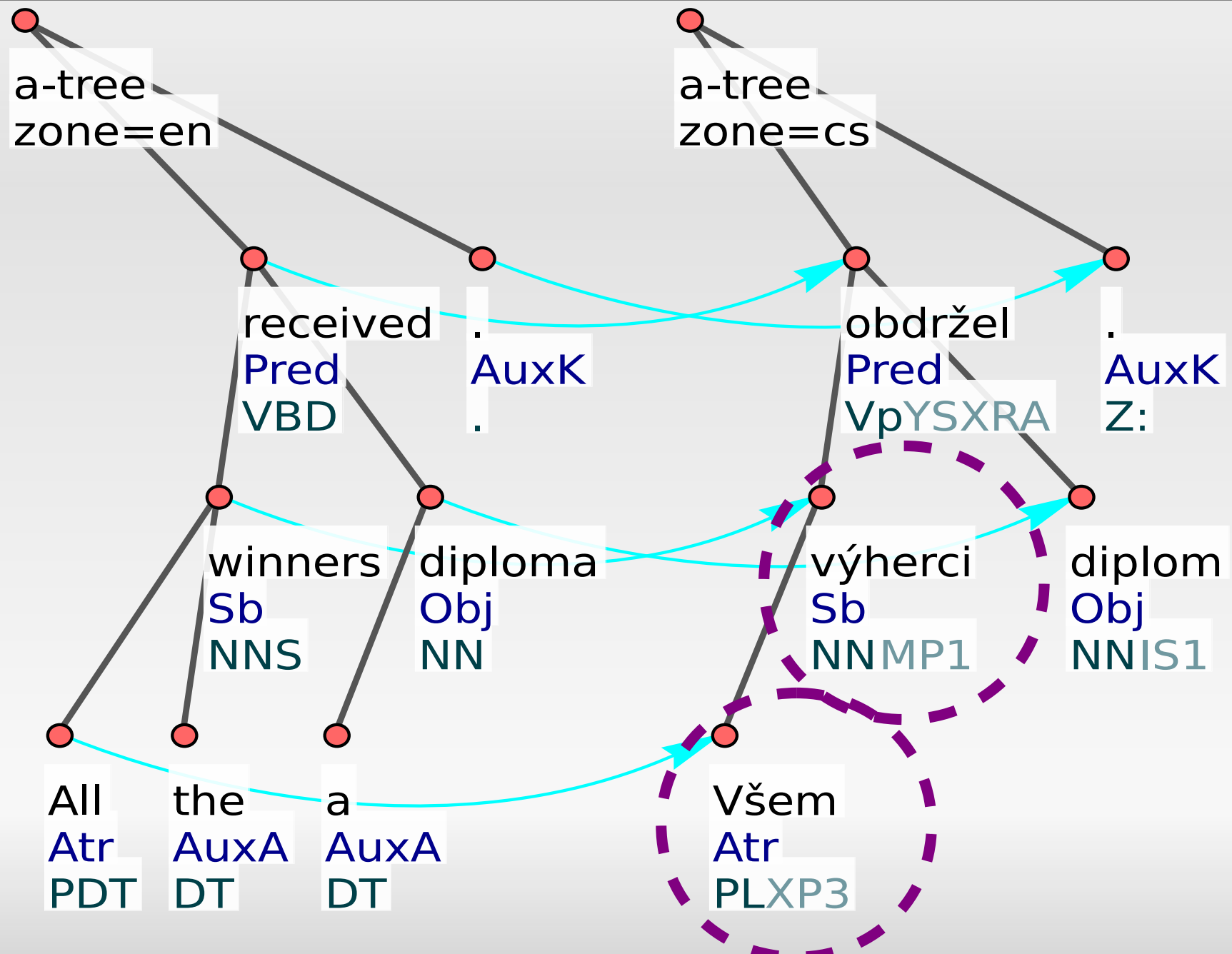
Podmět → nominativ



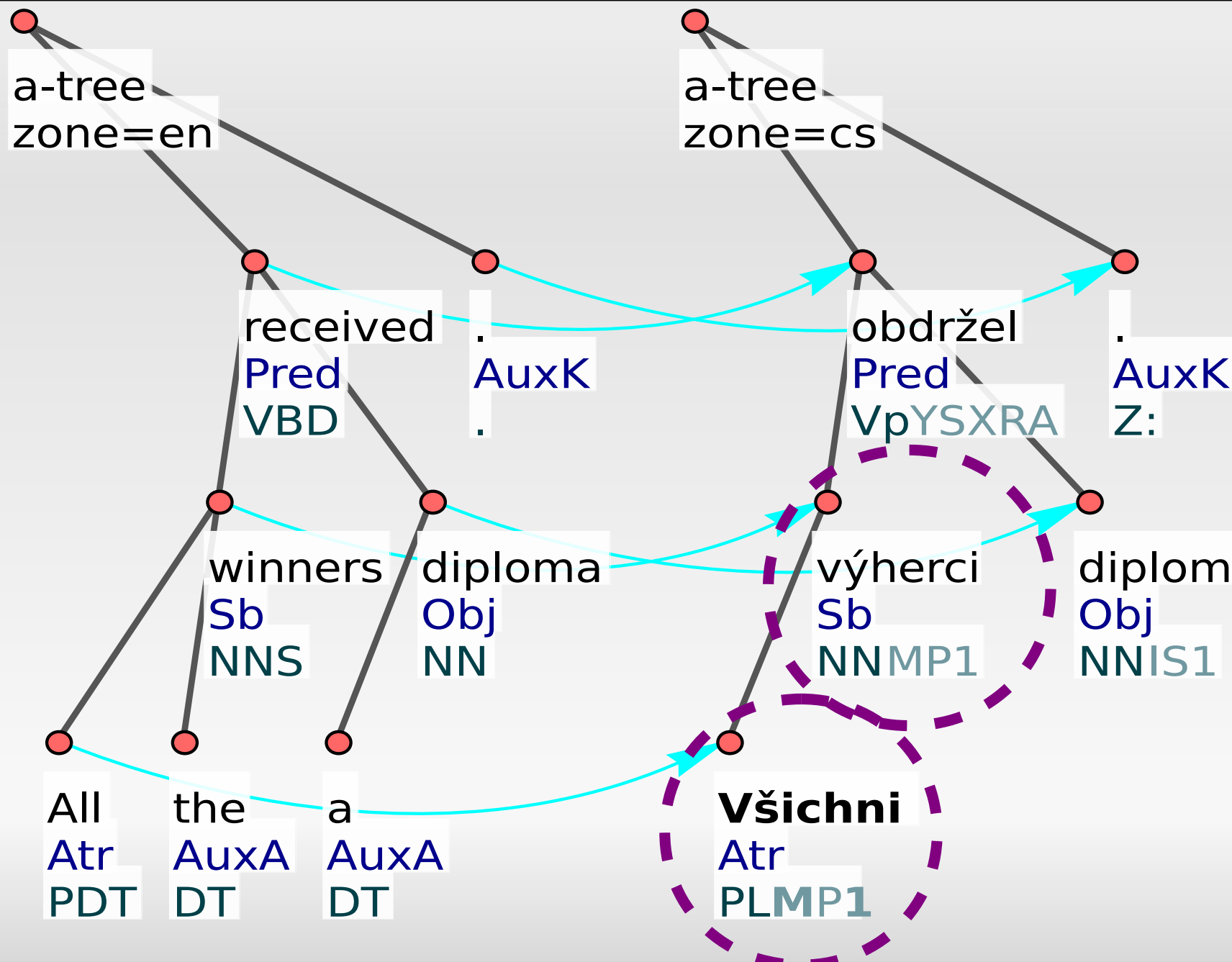
Všem výherci obdržel diplom.



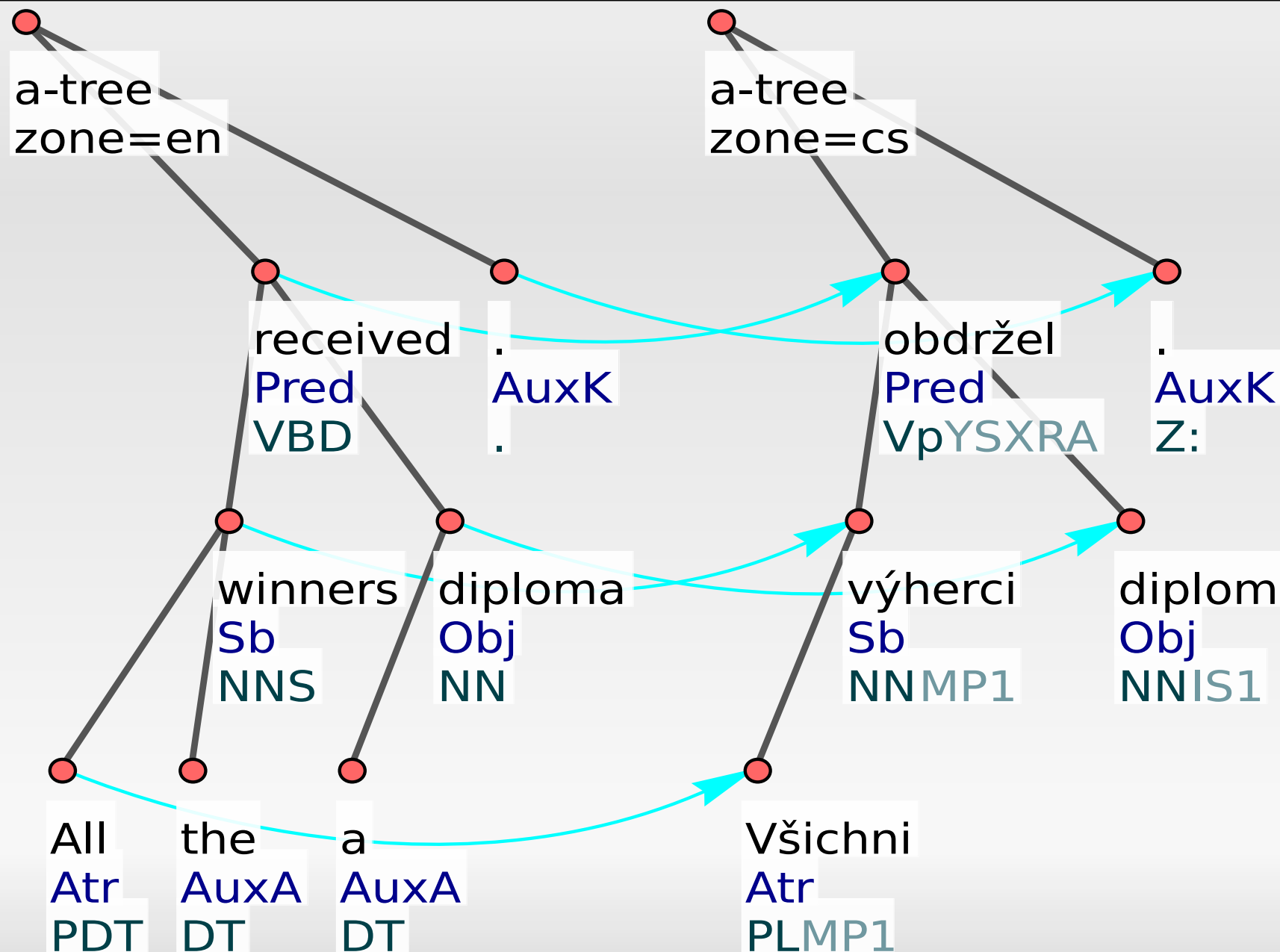
Shoda adjektiva se substantivem



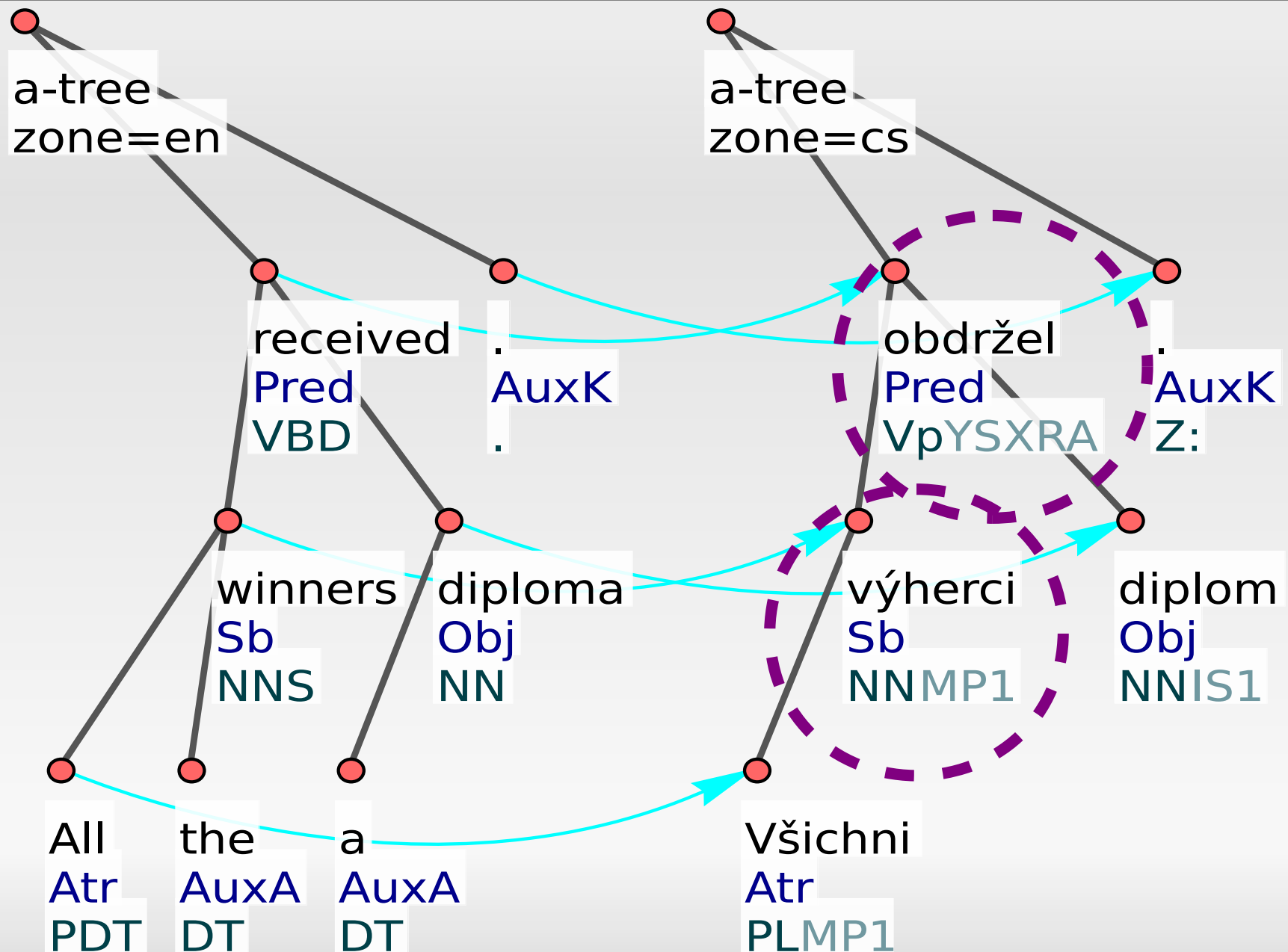
Shoda: rod, pád (číslo)



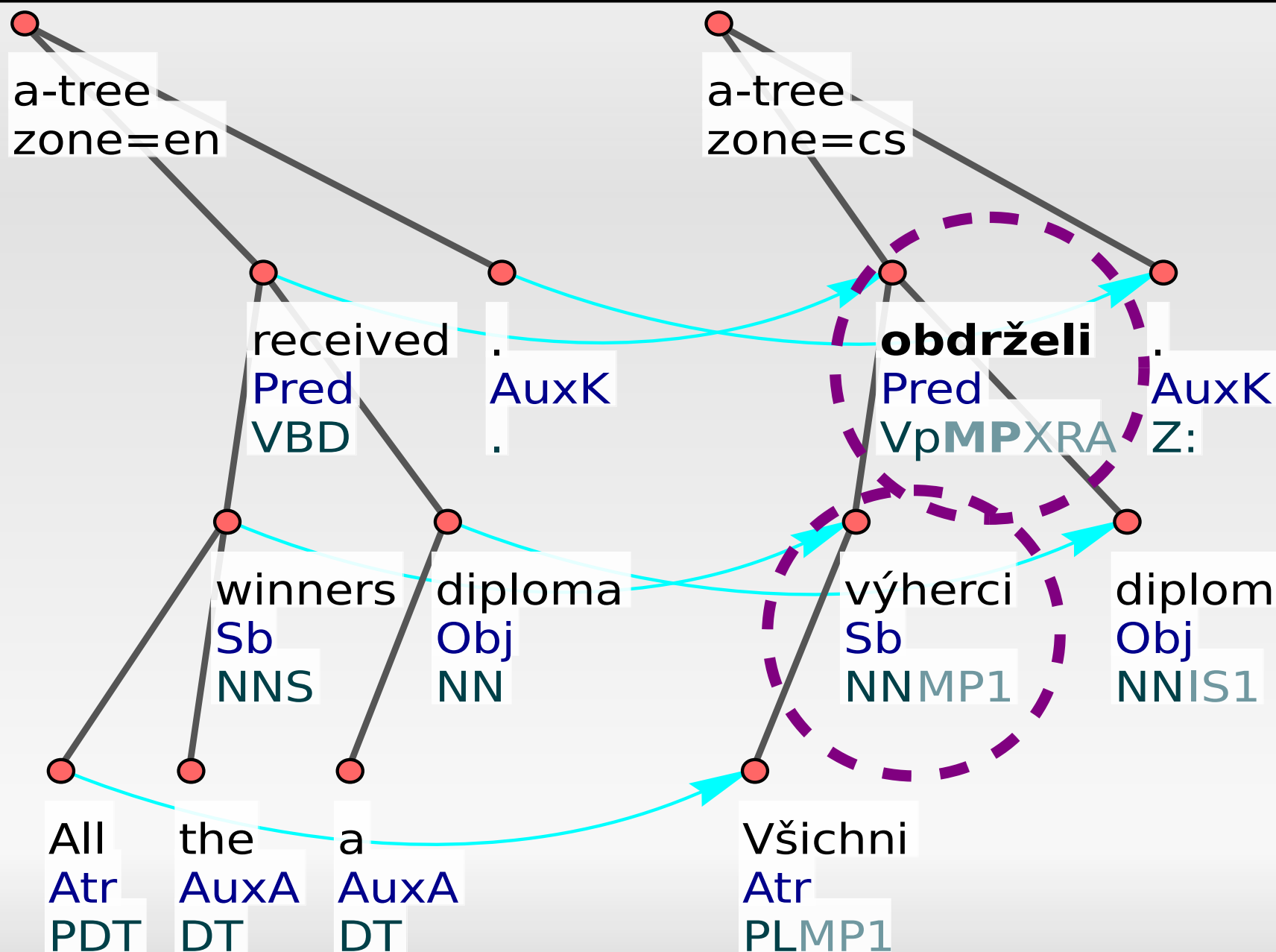
Všichni výherci obdržel diplom.



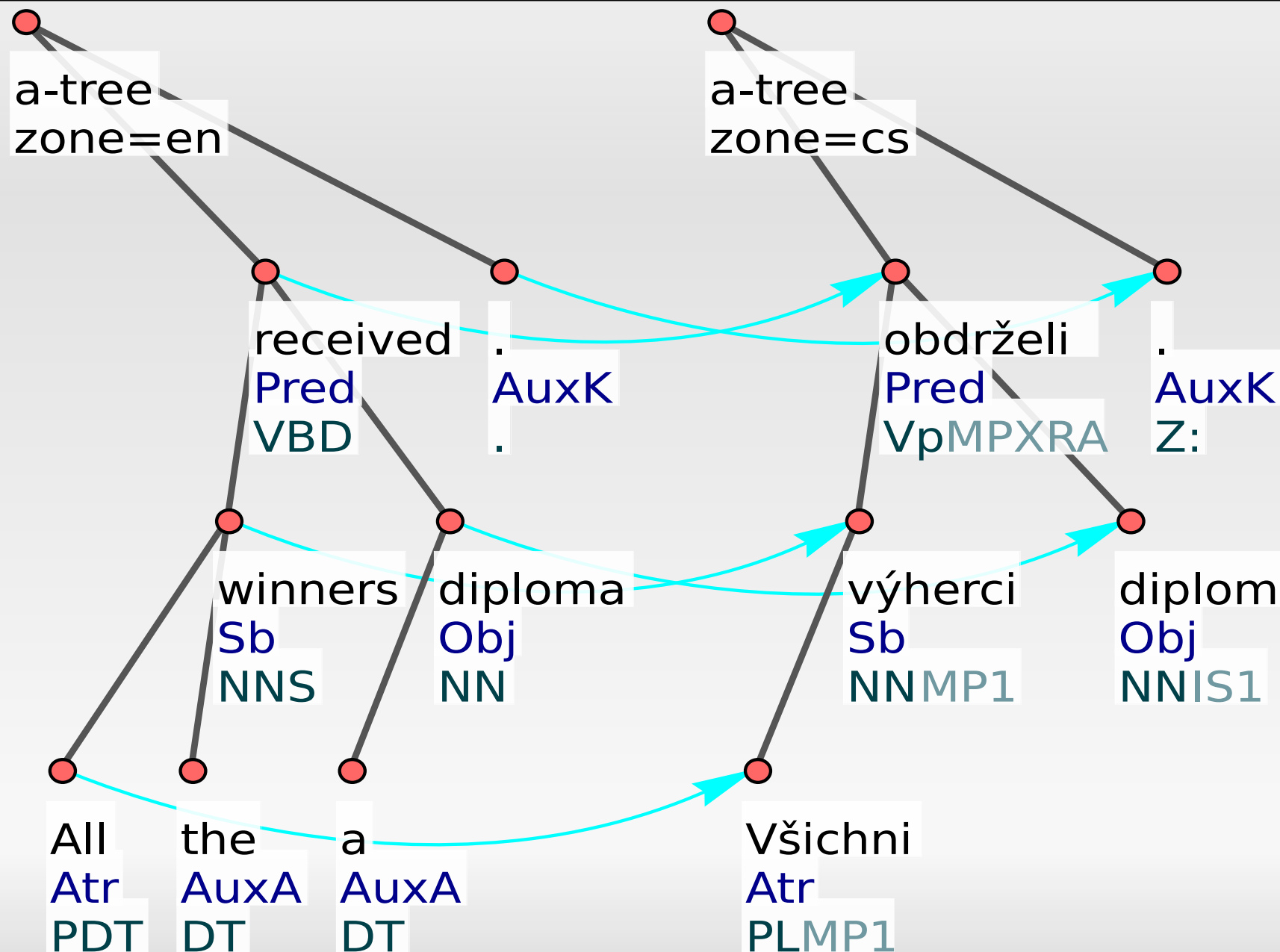
Shoda podmětu s přísudkem



Shoda: rod, číslo (osoba)



Všichni výherci obdrželi diplom.



T-rovina

- Analýza: t-stromy, formémy
- Opravy:
 - **pravidlové:** negace, překlad slovesných časů, vypouštění pronominálního podmětu
 - **statistické:** substantivní a verbální valence

Valenční model

- pravděpodobnost formému argumentu podmíněná
 - lematem rodiče (sloveso/substantivum)
 - formémem anglického argumentu
 - $\pm$ lematem argumentu
- trénováno na CzEngu

Úpravy použitých nástrojů

- nástroje pro automatickou analýzu jazyka
 - tagger, word-aligner, parser
- určené pro analýzu bezchybných vět
 - výstupy strojového překladu obsahují chyby
 - nástroje mají při jejich analýze nižší úspěšnost
- zvýšení robustnosti nástrojů
 - pravidlové opravy výstupů (tagger, aligner, parser)
 - reimplementace a modifikace parseru (pro češtinu)

Maximum Spanning Tree parser

- McDonald, Crammer, Pereira (2005)
 - Online large-margin training of dependency parsers

Maximum Spanning Tree parser

- McDonald, Crammer, Pereira (2005)
 - Online large-margin training of dependency parsers



Maximum Spanning Tree parser

- McDonald, Crammer, Pereira (2005)
 - Online large-margin training of dependency parsers
- McDonald, Pereira, Rybarov, Hajič (2006)
 - **Non-projective** dependency parsing using spanning tree algorithms



Maximum Spanning Tree parser

- reimplementace, vyladění pro češtinu



Maximum Spanning Tree parser

- reimplementace, vyladění pro češtinu
- „zhoršení“ trénovacích dat
 - zavlečení chyb podle chybového modelu, natrénovaného na výstupech Moses

Maximum Spanning Tree parser

- reimplementace, vyladění pro češtinu
- „zhoršení“ trénovacích dat
 - zavlečení chyb podle chybového modelu, natrénovaného na výstupech Moses
- přidání informací o zdrojové větě
 - tag, analytická funkce, existence hrany



Maximum Spanning Tree parser

- reimplementace, vyladění pro češtinu
- „zhoršení“ trénovacích dat
 - zavlečení chyb podle chybového modelu, natrénovaného na výstupech Moses
- přidání informací o zdrojové větě
 - tag, analytická funkce, existence hrany
- přidání informací z velkého korpusu (CzEng)
 - $PMI (parent, child) = \log \frac{p([parent, child])}{p([parent, *]) \cdot p([*, child])}$



Manuální vyhodnocení Depfixu

Vyhodnoceno	1350 vět
-------------	----------

Změněno	739 vět
---------	---------

Manuální vyhodnocení Depfixu

Vyhodnoceno	1350 vět
Změněno	739 vět
Zlepšeno	430 vět
Zhoršeno	152 vět
Neurčeno	157 vět

Manuální vyhodnocení Depfixu

Vyhodnoceno	1350 vět
-------------	----------

Změněno	739 vět
---------	---------

Zlepšeno	430 vět
----------	---------

Zhoršeno	152 vět
----------	---------

Neurčeno	157 vět
-----------------	---------

- správně → správně
- špatně → špatně
- věta zcela špatně, nelze určit

Manuální vyhodnocení Depfixu

Vyhodnoceno	1350 vět
Změněno	739 vět
Zlepšeno	430 vět
Zhoršeno	152 vět
Neurčeno	157 vět
„Úspěšnost“	58,2%
„Přesnost“	73,9%
„Úplnost“	31,9%

$$\text{„úspěšnost“} = \frac{\text{zlepšeno}}{\text{změněno}}$$

$$\text{„přesnost“} = \frac{\text{zlepšeno}}{\text{zlepšeno} + \text{zhoršeno}}$$

$$\text{„úplnost“} = \frac{\text{zlepšeno}}{\text{vyhodnoceno}}$$

Automatické vyhodnocení (BLEU)

Systém	WMT 2011		WMT 2012	
	před	po	před	po
Joshua (Zeman)	14, <u>0</u> 8	14, <u>8</u> 1	12, <u>1</u> 0	12, <u>4</u> 4
Moses (Bojar)	16, <u>3</u> 5	16, <u>8</u> 3	14, <u>1</u> 9	14, <u>2</u> 6
Moses (Koehn)	17, <u>3</u> 0	17, <u>9</u> 4	15, <u>5</u> 4	15, <u>7</u> 8
Google Translate	19, <u>7</u> 3	19, <u>9</u> 7	16,22	16,22
... a dalších 9 systémů ...				
průměrné zlepšení	+0,38		+0,19	

Shrnutí

- Depfix = automatická post-editace výstupů frázového strojového překladu (AJ → ČJ)
- Lingvistická analýza v Treexu (M-, A-, T-rovina)
 - zvýšena robustnost nástrojů (data obsahují chyby)
- Pravidlové i statistické opravy
 - shody, negace, podmínky, časy, valence...
- Zvýšení kvality strojového překladu
 - do značné míry nezávislé na překladovém systému

Děkuji za pozornost

- Rudolf Rosa, ÚFAL MFF UK
 - rosa@ufal.mff.cuni.cz
 - <http://ufal.mff.cuni.cz/~rosa/>
 - tato prezentace i samotná diplomová práce budou k dispozici na webu
- Depfix je součástí Treexu
 - trunk/treex/devel/depfix/
 - free software; můžete jej redistribuovat a/nebo modifikovat za stejných podmínek jako Perl

Generování

- většina oprav vede ke změně tagu jednoho nebo více slov
- pro nalezení správné slovní formy používáme Hajičův morfologický generátor
- `forma = generátor(lemma, tag)`