Rudolf Rosa

diplomová práce

Automatická post-editace výstupů frázového strojového překladu (Depfix)

Automatic post-editing
of phrase-based machine translation outputs

Motivační příklad

- Zdroj:
  - *All the winners received a diploma.*

Motivační příklad

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

Motivační příklad

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

Motivační příklad

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

Motivační příklad

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

Motivační příklad

- Zdroj:
  - ***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:
  - *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:
  - *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
- Vyhodnocení

Depfix krok za krokem

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

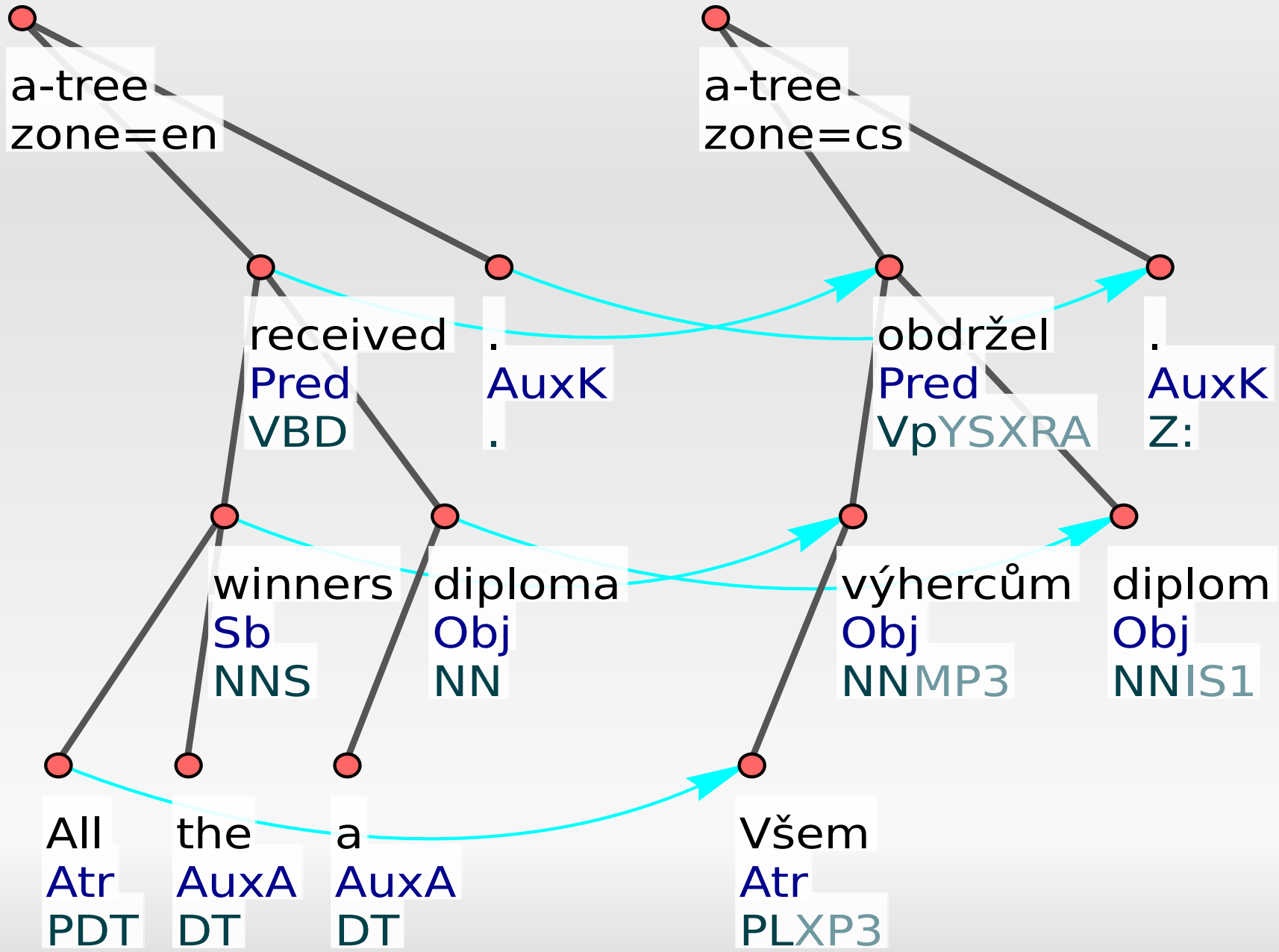
M-rovina

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

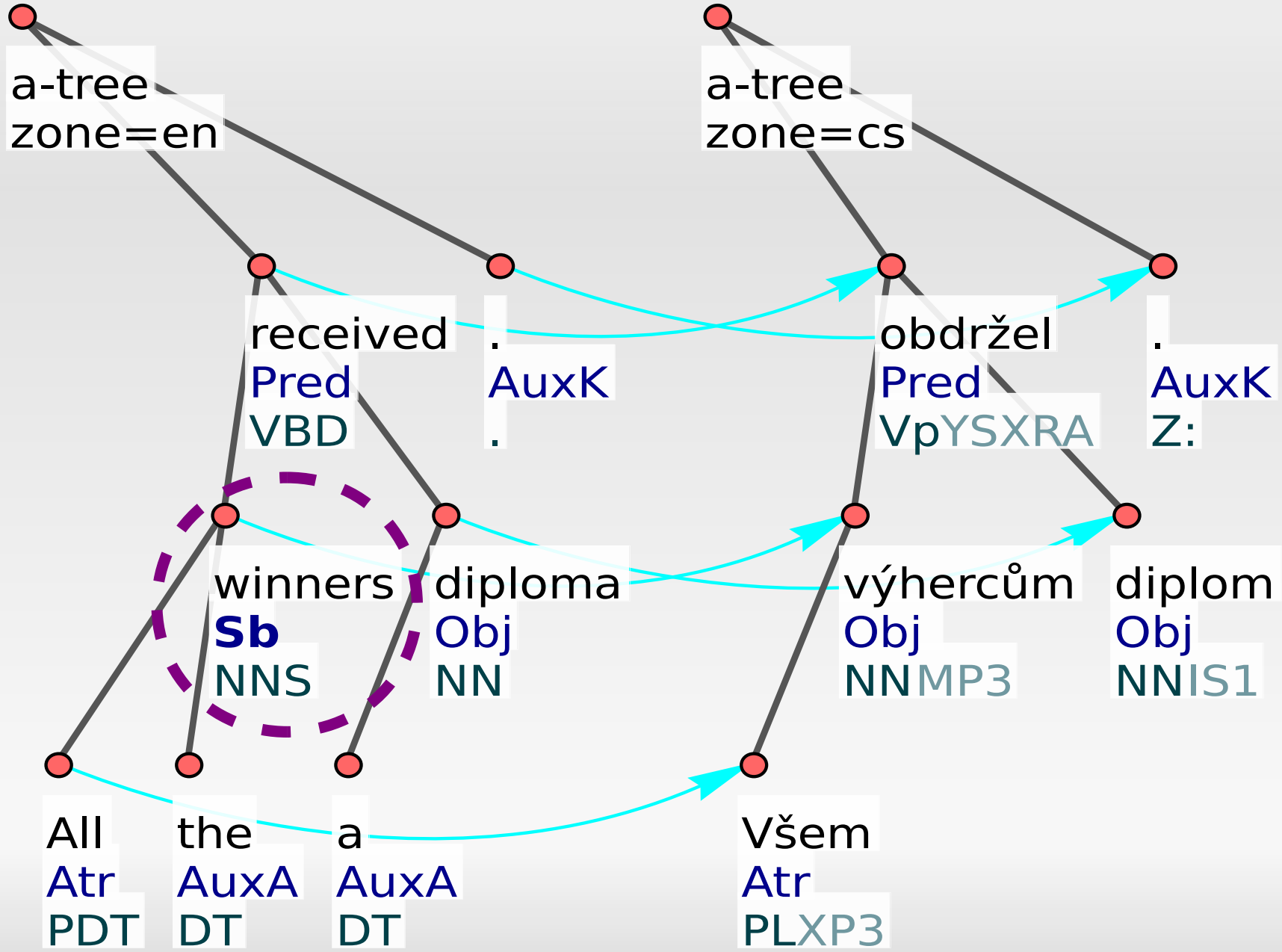
A-rovina

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

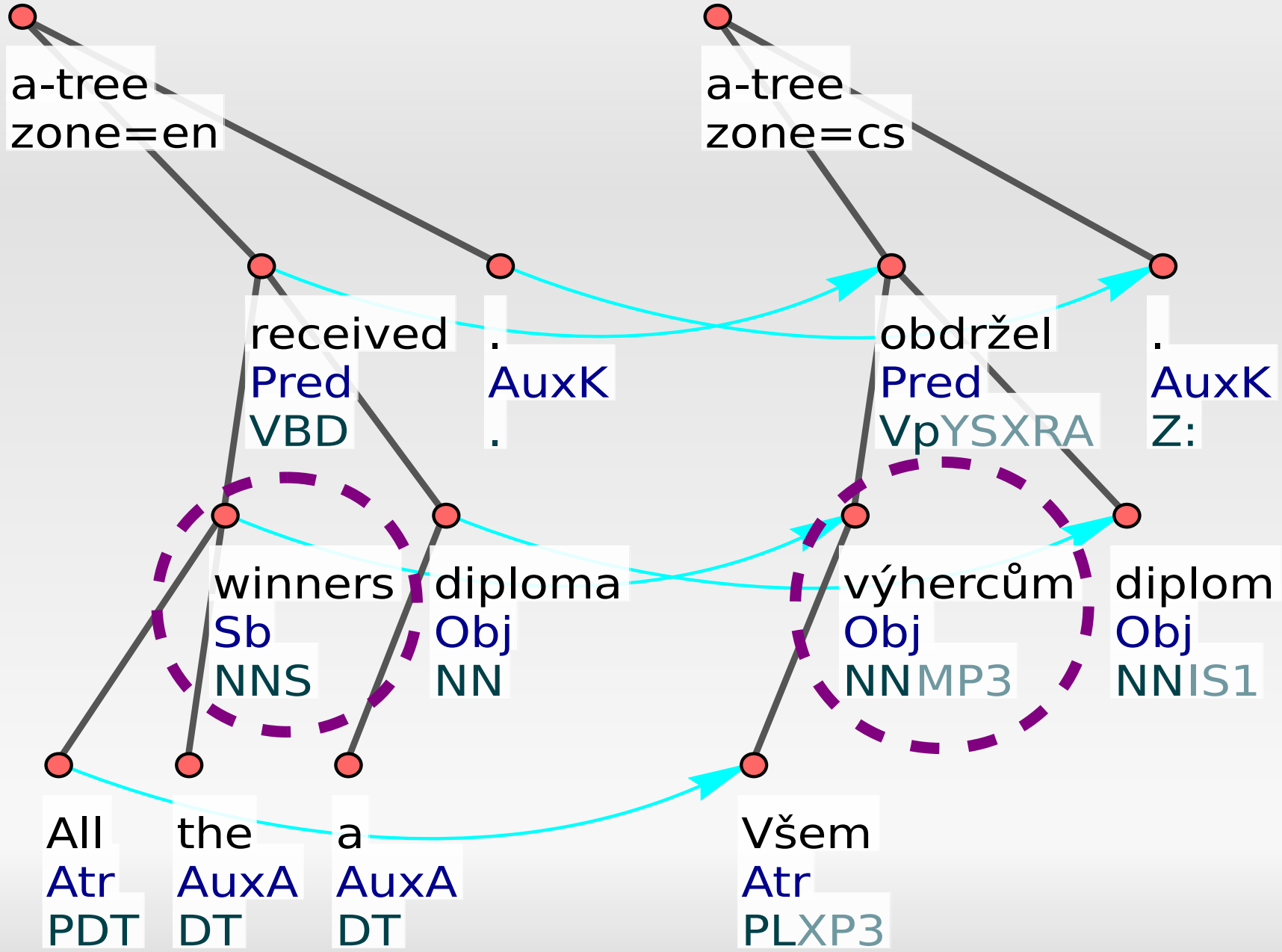
Všem výhercům obdržel diplom.



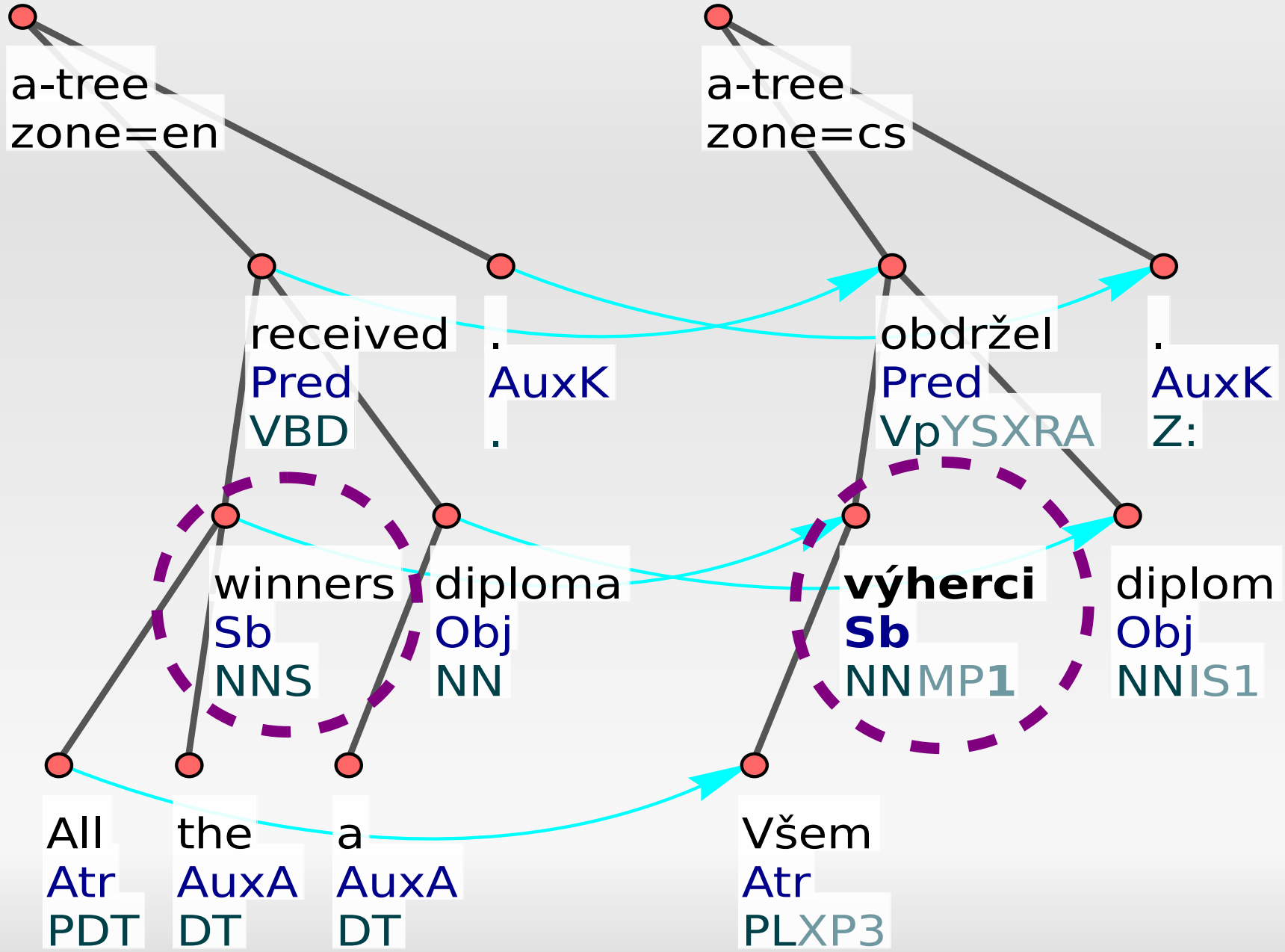
Transfer významu: podmět



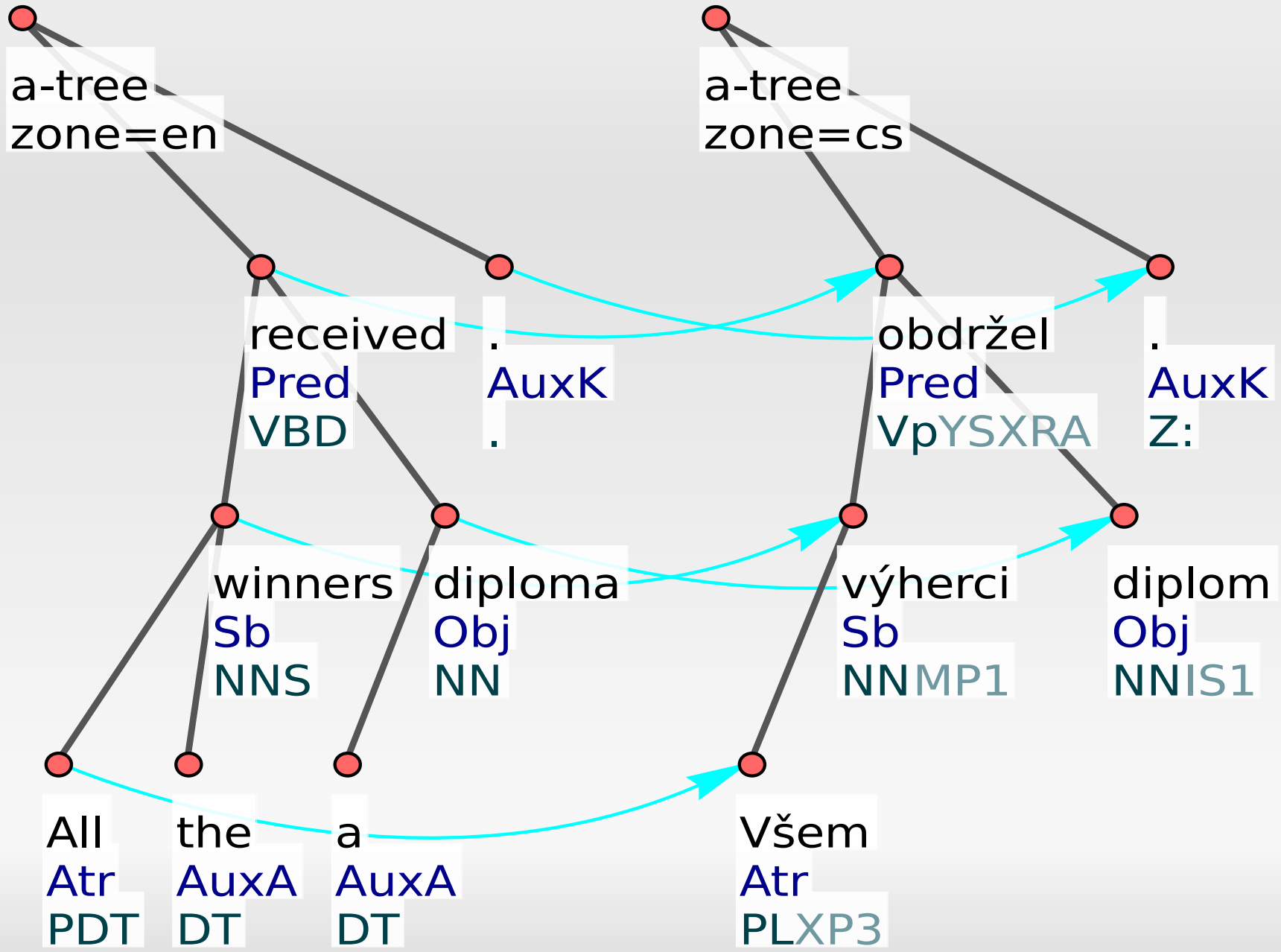
Transfer významu: podmět



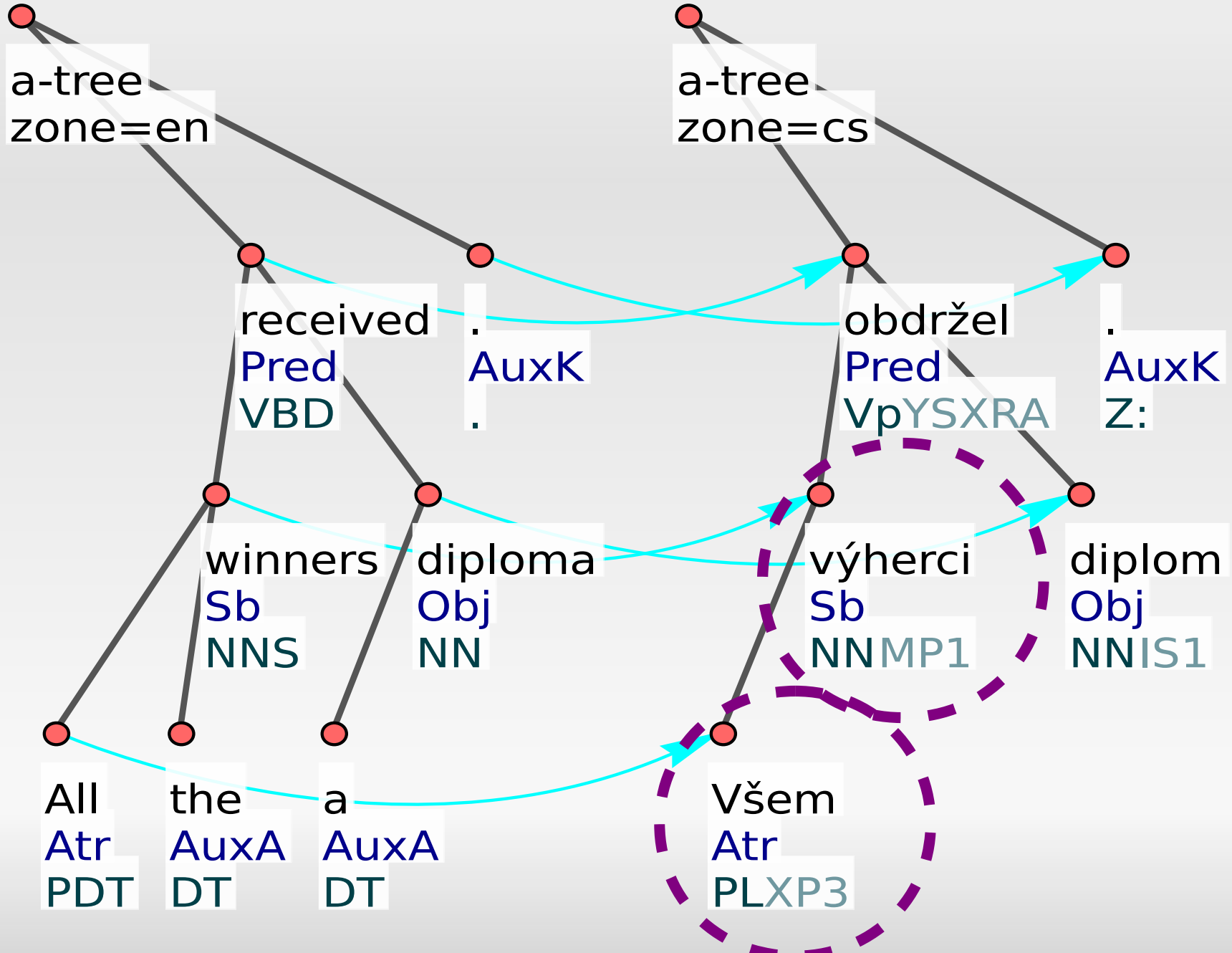
Podmět → nominativ



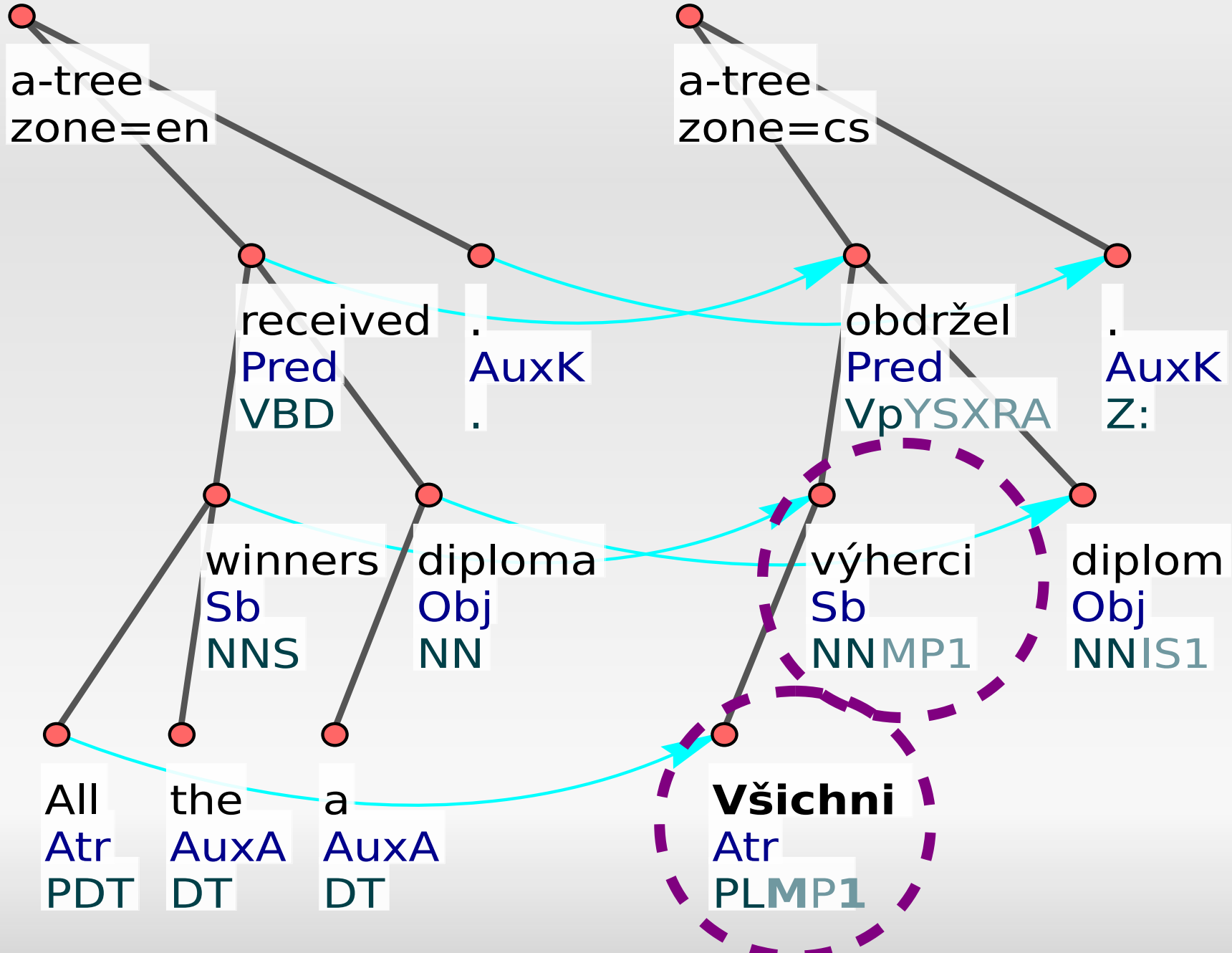
Všem výherci obdržel diplom.



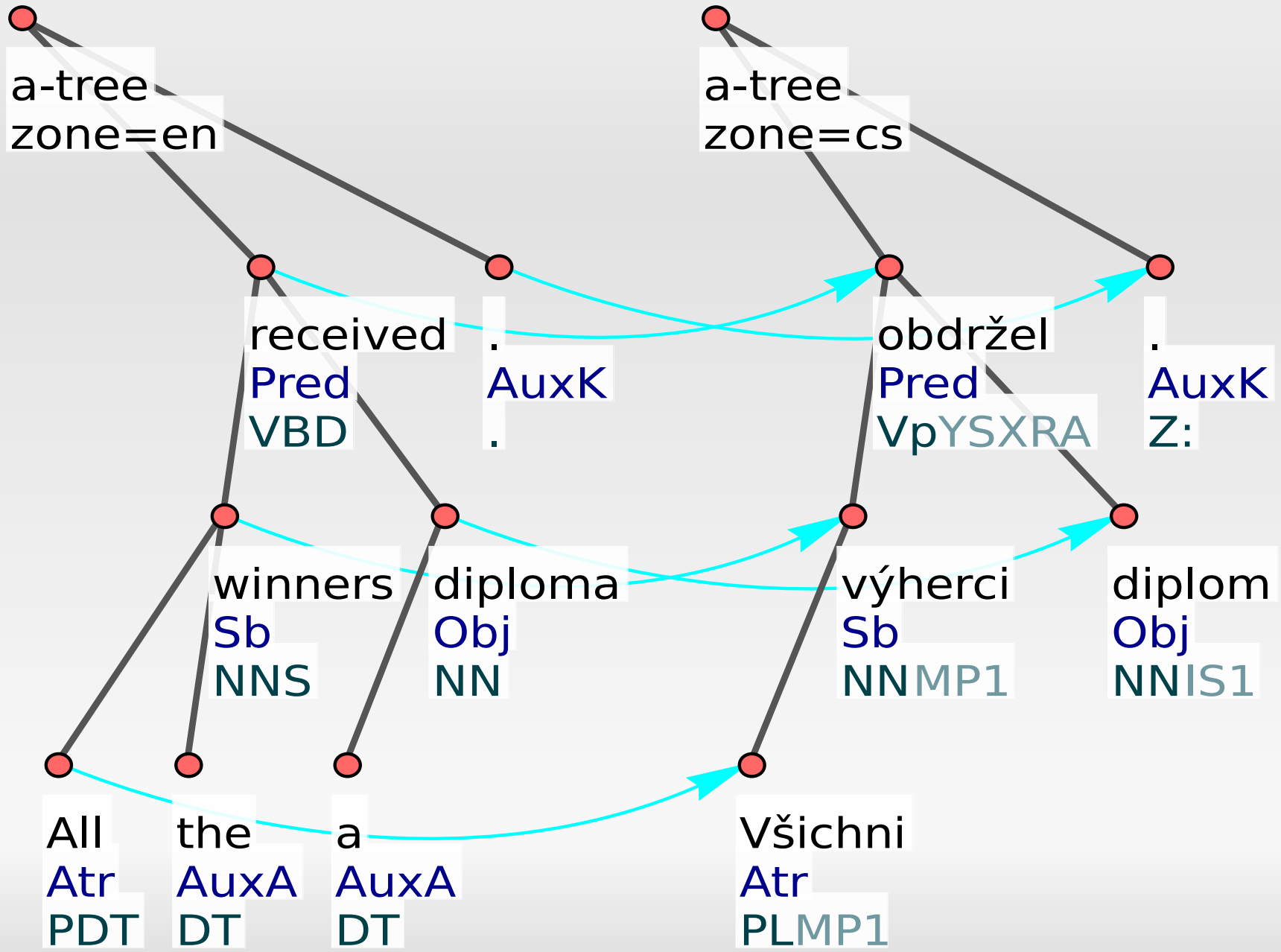
Shoda adjektiva se substantivem



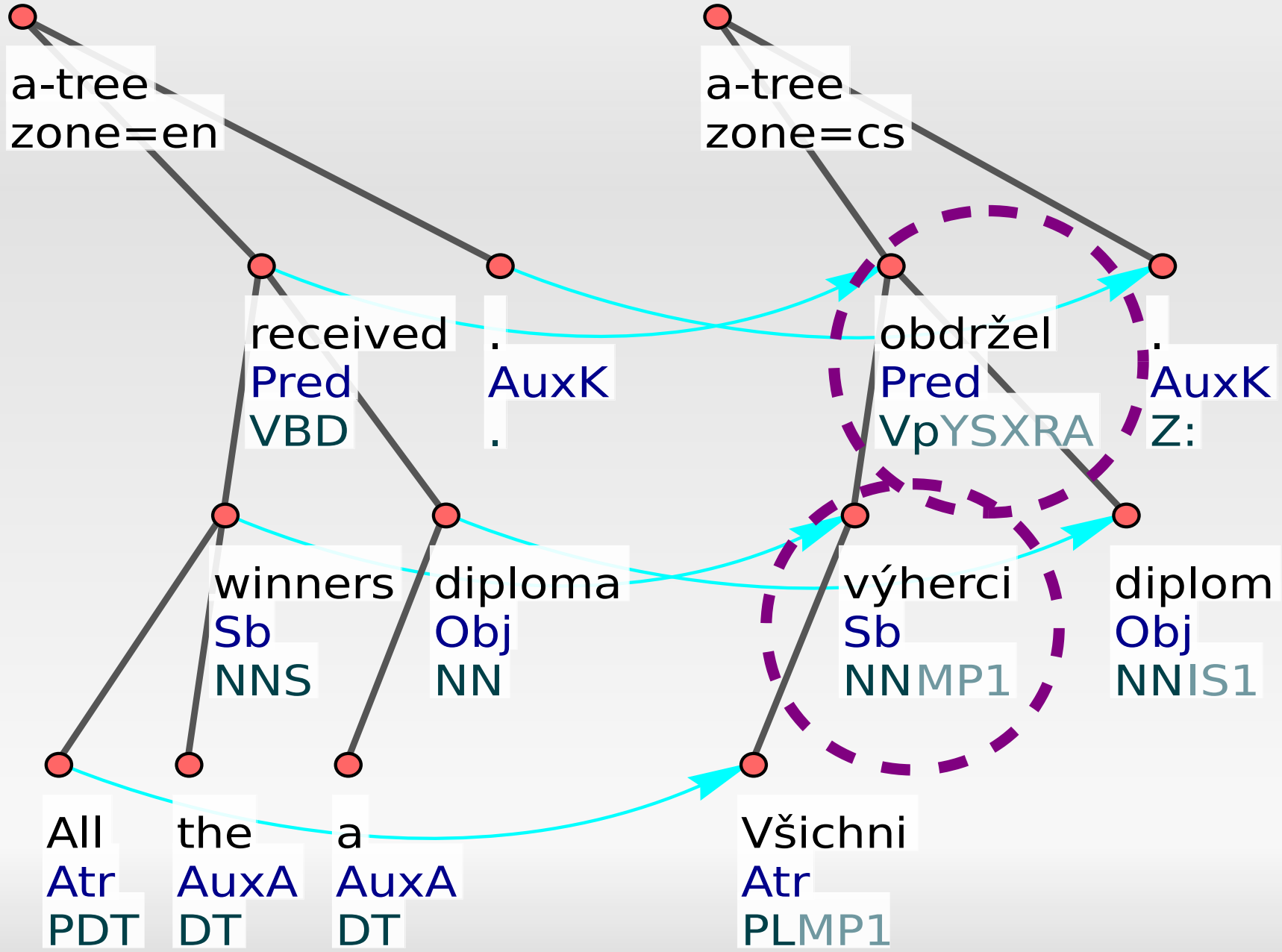
Shoda: rod, pád (číslo)



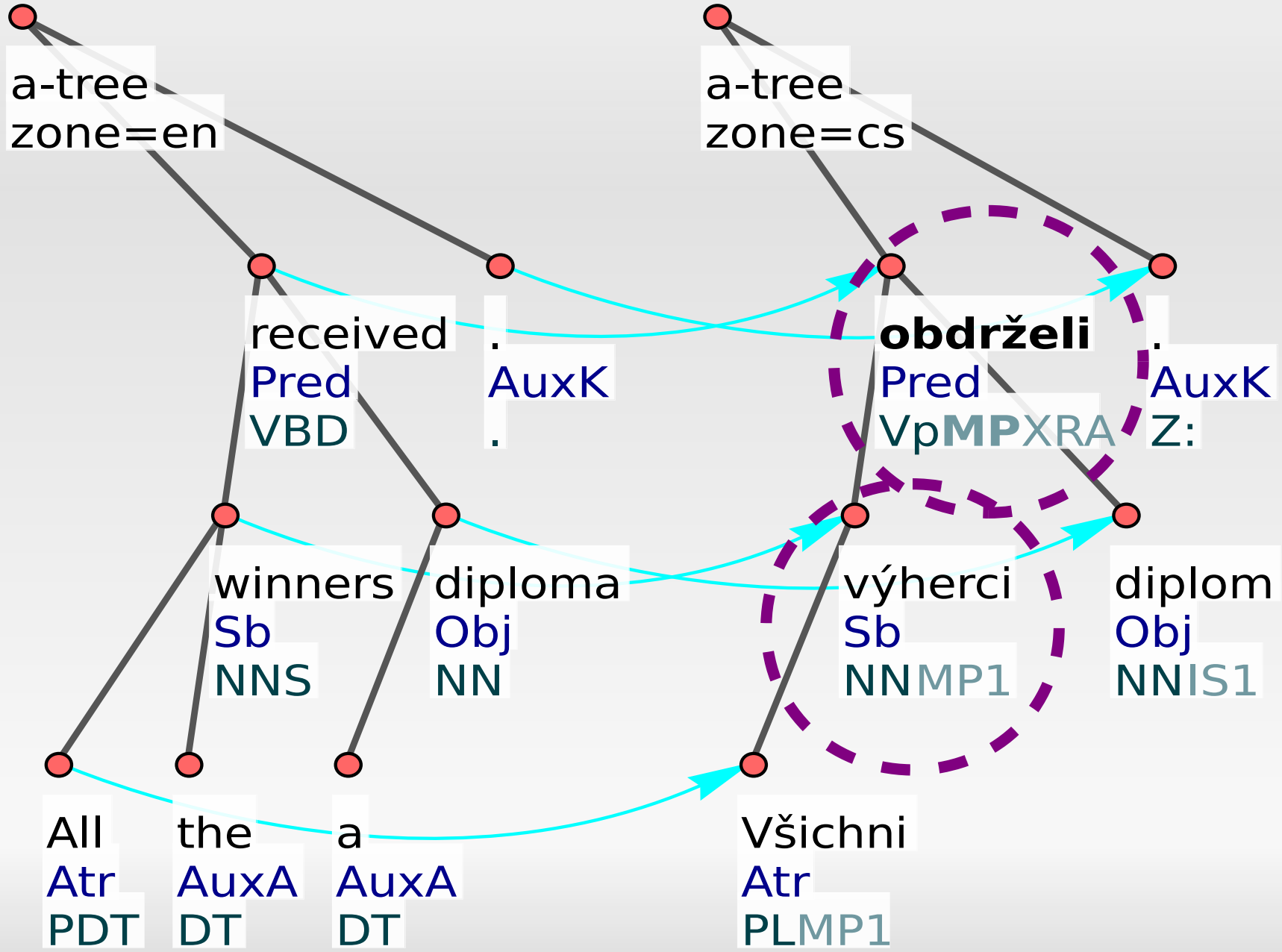
Všichni výherci obdržel diplom.



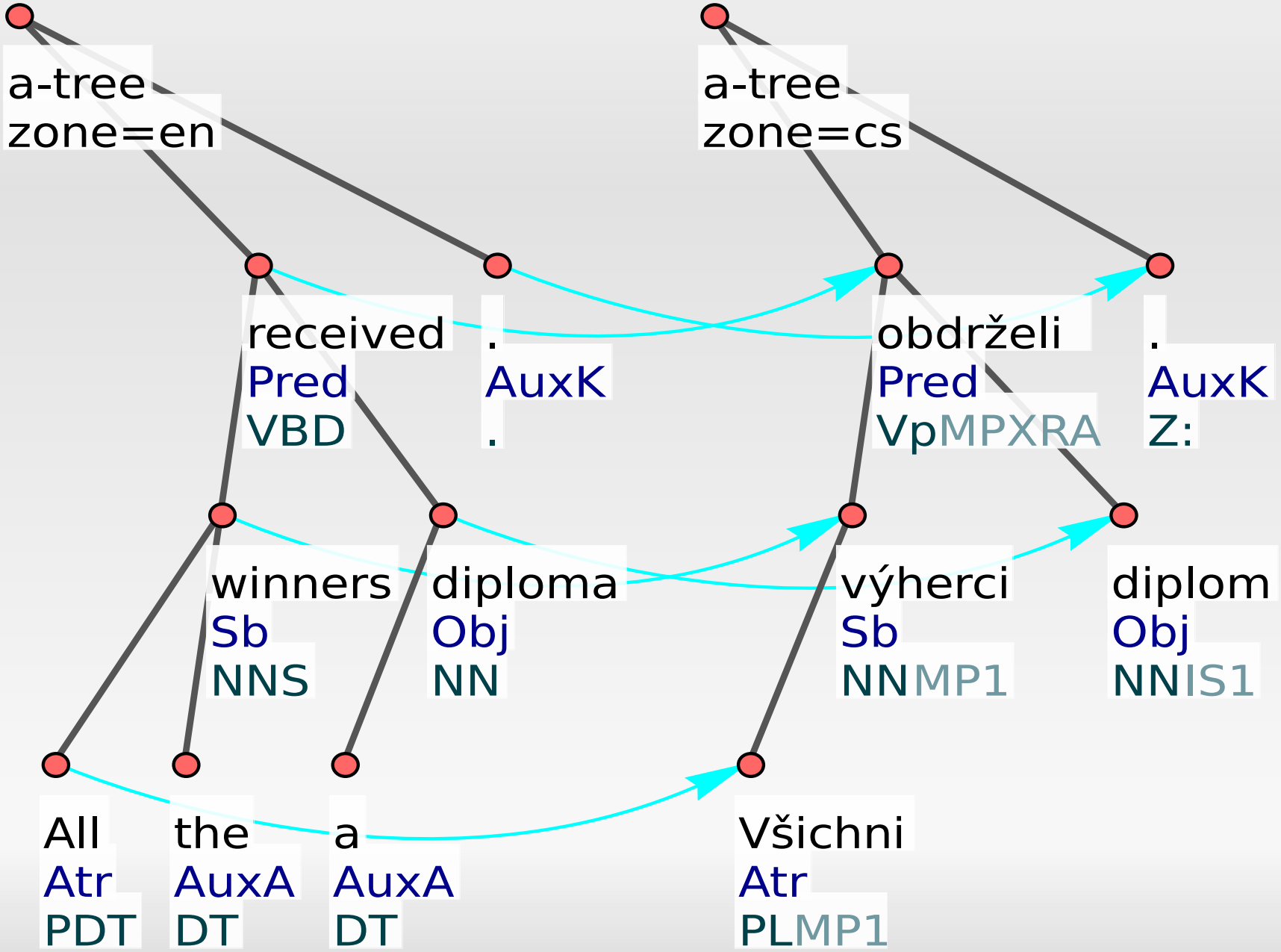
Shoda podmětu s přísudkem



Shoda: rod, číslo (osoba)



Všichni výherci obdrželi diplom.



A-rovina

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

T-rovina

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

Valenční model

- natrénováno na CzEngu
- pravděpodobnost formému argumentu podmíněná
  - lemmatem rodiče (sloveso/substantivum)
  - formémem anglického argumentu
  - ± lemmatem argumentu

Úpravy použitých nástrojů

- nástroje pro automatickou analýzu jazyka
  - tagger, word-aligner, parser
- určené pro analýzu bezchybných vět
  - výstupy strojového překladu obsahují chyby
  - nátroje 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 Mosese

Maximum Spanning Tree parser

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

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
„Úspěšnost“	58,2%
„Přesnost“	73,9%
„Úplnost“	31,9%

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

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

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

Automatické vyhodnocení (BLEU)

Systém	WMT 2011		WMT 2012	
	před	po	před	po
Joshua (Zeman)	14,08	14,81	12,10	12,44
Moses (Bojar)	16,35	16,83	14,19	14,26
Moses (Koehn)	17,30	17,94	15,54	15,78
Google Translate	19,73	19,97	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, podměty, 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 dipozici na webu
- Depfix je součástí Treexu
  - trunk/treex/devel/depfix/
  - free software; můžete jej redistribuovat a/nebo modifikovat za stejných podmínek jako Perl