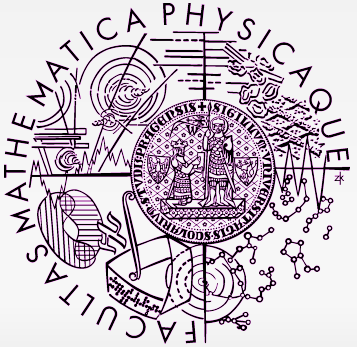
Rudolf Rosa
rur@nikde.eu
http://ufal.mff.cuni.cz/rudolf-rosa

Depfix:

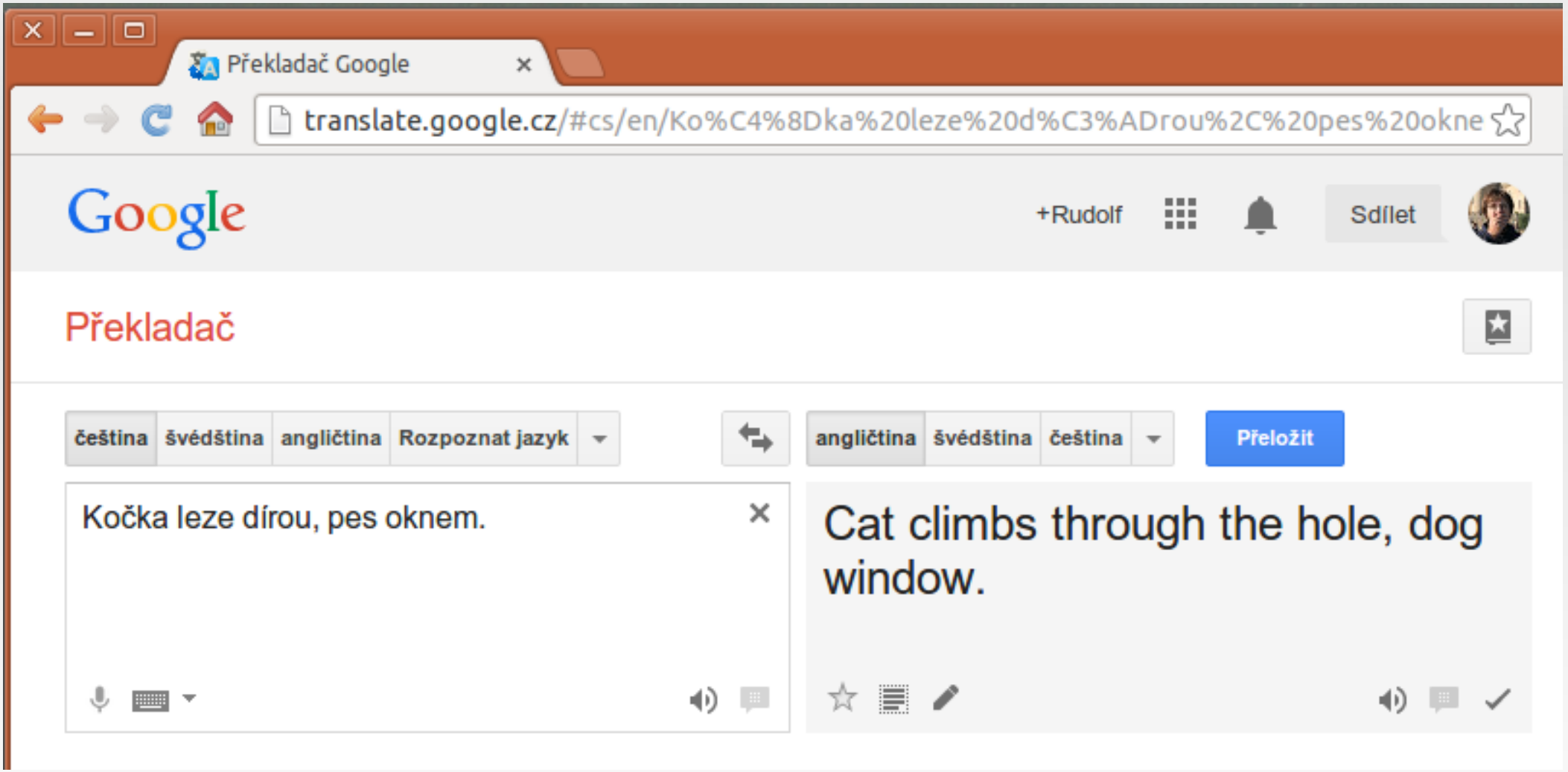
Jak dělat strojový překlad lépe než Google Translate



Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

ČLO, Praha, 22. dubna 2014

Jak překládá strojový překladač?



Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

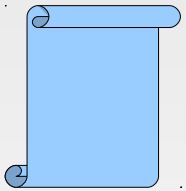
- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- ???

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- …

- ptakopysk plave ve vodě
- ježura žije v austrálii
- …



paralelní korpus

- česko-anglický paralelní korpus (CzEng): 15 000 000 vět
  - knihy, legislativa EU, titulky, dvojjazyčné weby...
- hindsko-anglický (HindEnCorp): 287 000 vět

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- ???

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- ???

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- plave

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- plave ???

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- plave tady

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- plave tady ???

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- cvync fala e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- plave tady ryba

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

- + pravděpodobnosti frázových párů
  - ym = ve (66%) / v (33%)

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

- rywac ym e brediboc e hycf

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

- rywac ym e brediboc e hycf ym = ve (66%) / v (33%)
- žije

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

- rywac ym e brediboc e hycf
- žije ve

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

- rywac ym e brediboc e hycf
- žije ve ptakopysk

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

- rywac ym e brediboc e hycf
- žije ve ptakopysk ryba

Jak překládá strojový překladač?

- cvync ym vedal e brediboc
- rywac ym eocdlerye em asfytme
- rywac fala e brediboc
- rywac ym vedal e hycf

- ptakopysk plave ve vodě
- ježura žije v austrálii
- ptakopysk žije tady
- ryba žije ve vodě

- cvync fala e hycf

- plave tady ryba

= frázový statistický strojový překlad (Google, Moses...)

- + změna slovosledu, + jazykový model:
- žije ve ptakopysk ryba → ryba žije v ptakopysk

Co frázový překlad (moc) neumí

- anglická věta:
  - *All the winners received a diploma.*
- překlad pomocí systému Moses:
  - *Všem výhercům obdržel diplom.*

- frázový překlad moc neumí gramatiku
  - neví, že něco je podmět, a že to má být v 1. pádě
  - neví, že slova mají číslo, a že by se mělo shodovat číslo podmětu a přísudku
  - …

Co frázový překlad (moc) neumí

- někdy se hodí dívat se „dovnitř slov“
  - skloňování (pád, číslo, rod..., negace...)
  - časování (čas, osoba, slovesný rod...)
- někdy se hodí dívat se na strukturu věty
  - shody (podmět-přísudek, jméno-přívlastek...)
  - slovesná valence (např. *někdo*$_1$ *poslal* *něco*$_4$ *někomu*$_3$)
- zapojit gramatiku přímo do frázového překladu je složité (byť to trochu jde)
- opravovat chyby zvlášť je jednodušší

Depfix – automatická post-editace

I. frázový strojový překlad

II. automatická oprava chyb

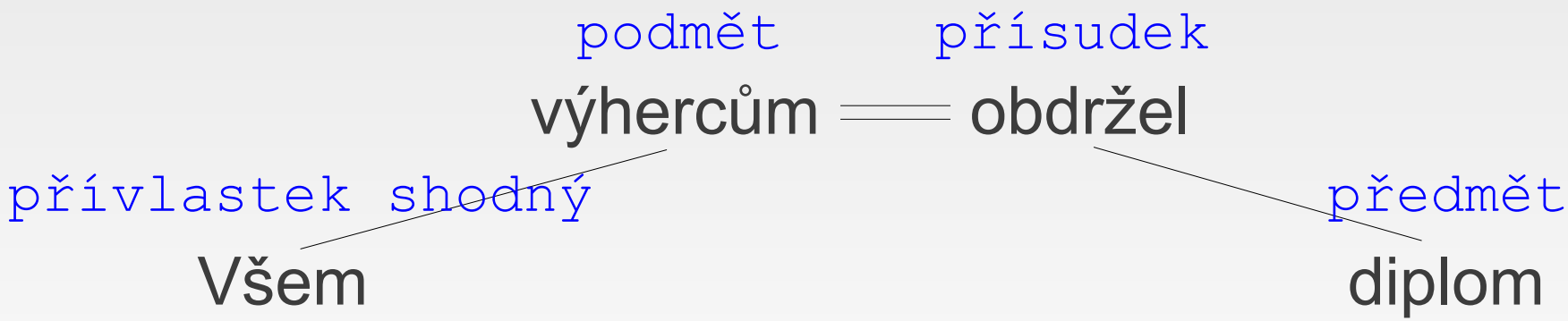
1. automatický jazykový rozbor
    - slovní druhy, mluvnické kategorie (98%)
    - větný rozbor, větné členy (80%)
2. automatická oprava chyb (Depfix)
    - pravidla (shoda, slovesný čas...)
    - statistika (slovesná valence)

Jak Depfix opravuje chyby

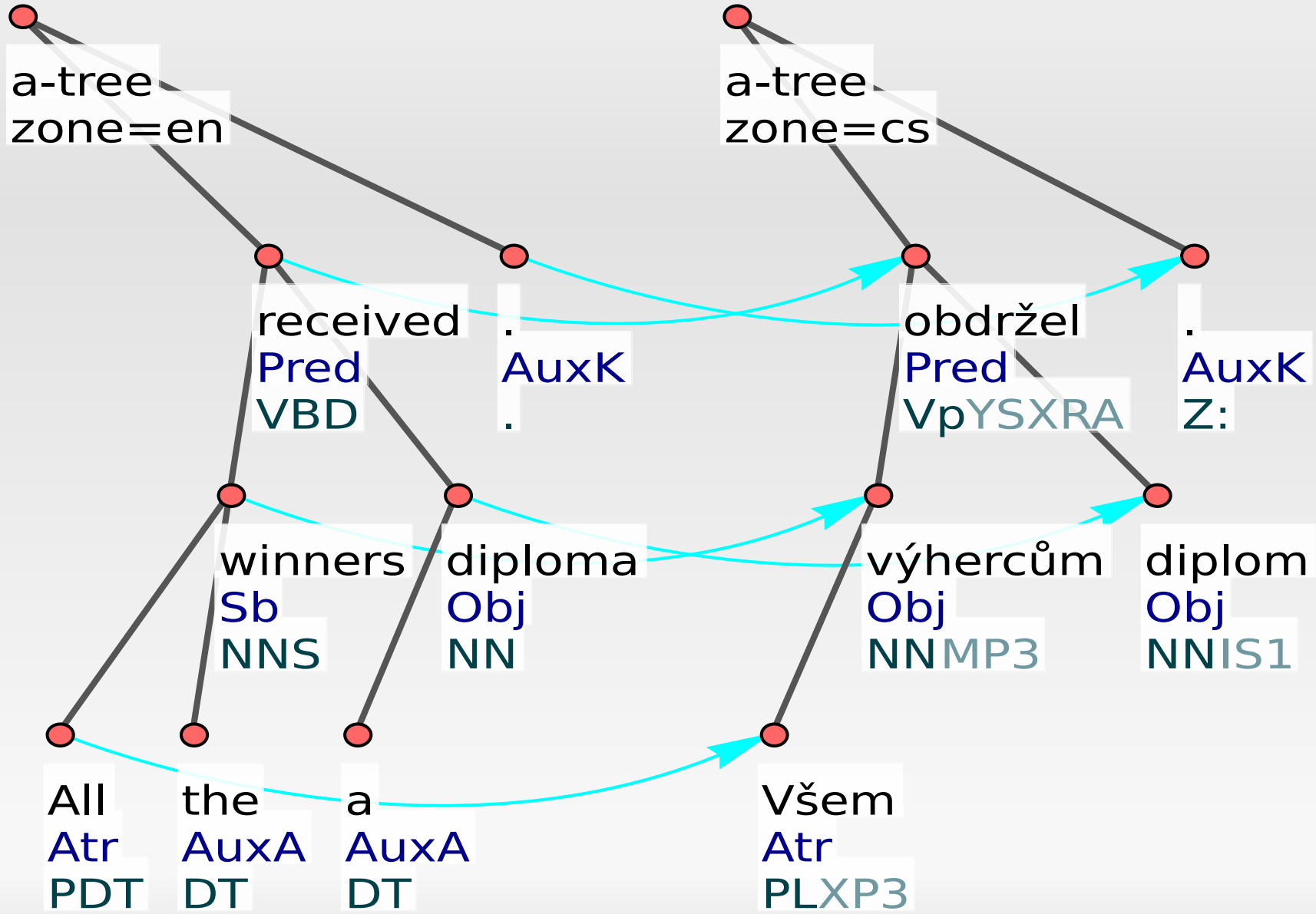
- anglická věta:
  - *All the winners received a diploma.*
- překlad pomocí systému Moses:
  - *Všem výhercům obdržel diplom.*
- oprava pomocí systému Depfix:
  - *Všichni výherci obdrželi diplom.*

Jazykový rozbor věty

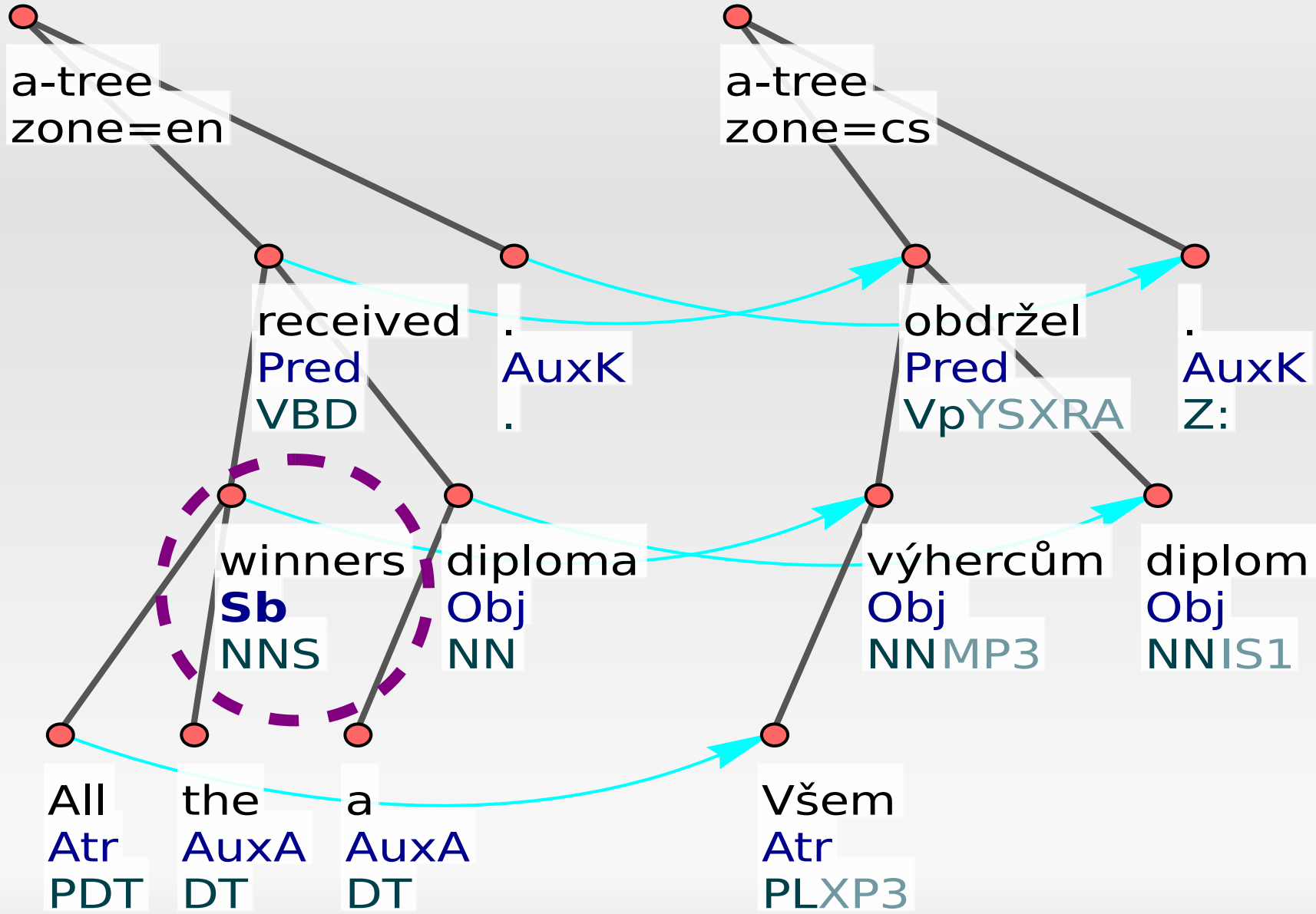
- slovní druhy, mluvnické kategorie
  - *Všem **výhercům** obdržel diplom.*
  - lemma *výherce*, podst. jm., 3. p., mn. č., rod m. živ.
- větný rozbor, větné členy



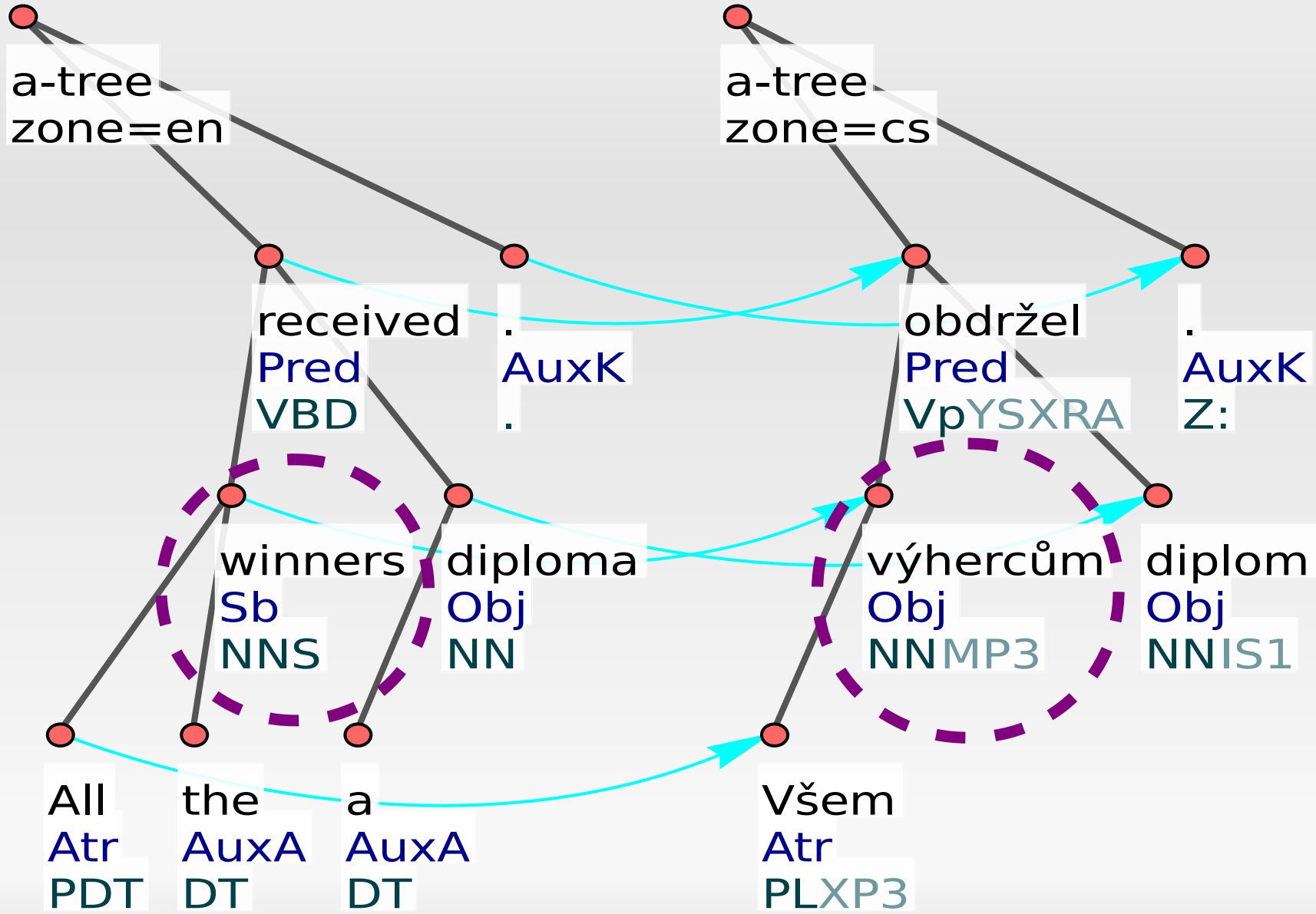
Všem výhercům obdržel diplom.



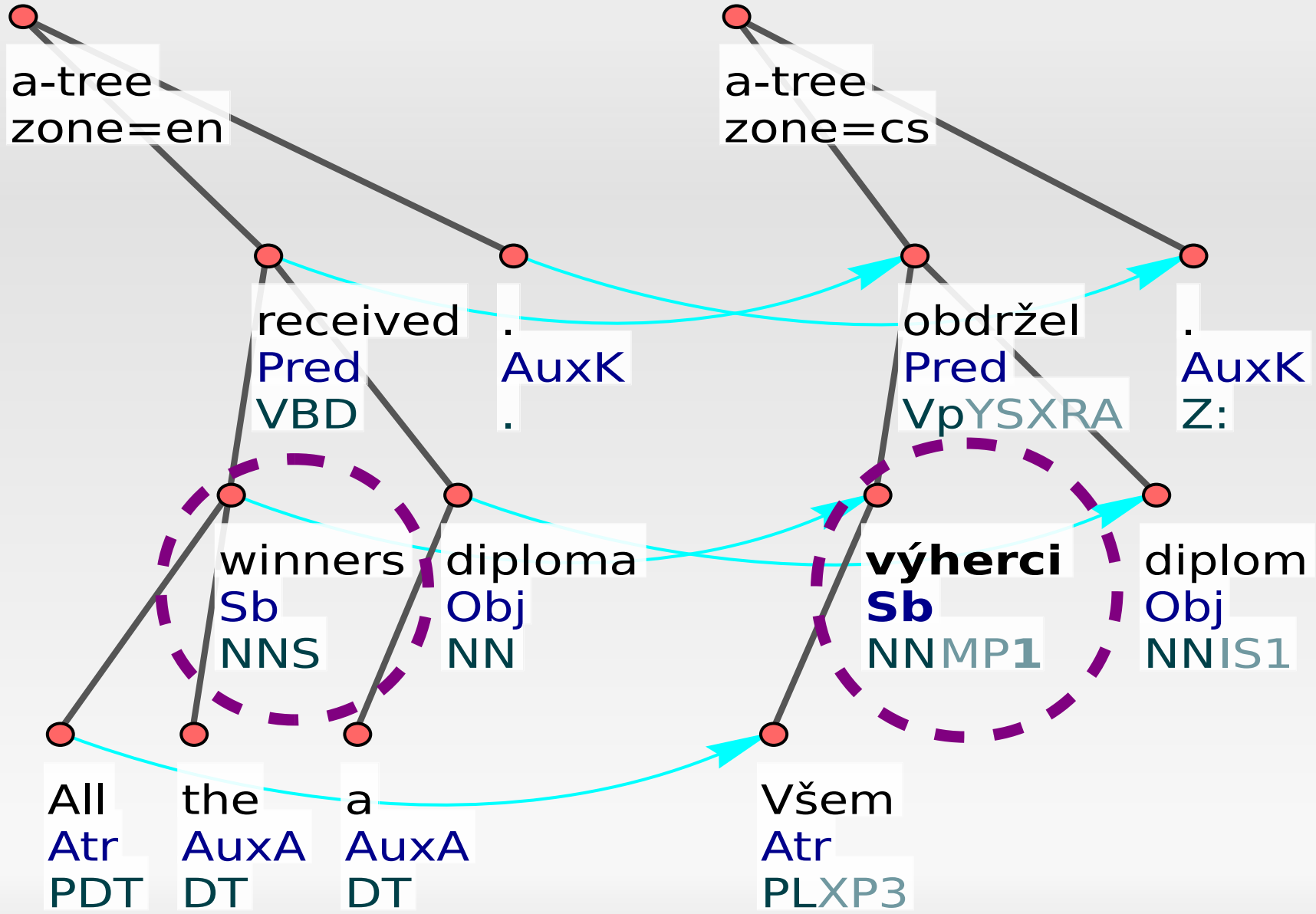
Podmět


a-tree
zone=en
a-tree
zone=cs
received
Pred
VBD
.
AuxK
.
obdržel
Pred
VpYSXRA
.
AuxK
Z:
winners
Sb
NNS
diploma
Obj
NN
výhercům
Obj
NNMP3
diplom
Obj
NNIS1
All
Atr
PDT
the
AuxA
DT
a
AuxA
DT
Všem
Atr
PLXP3

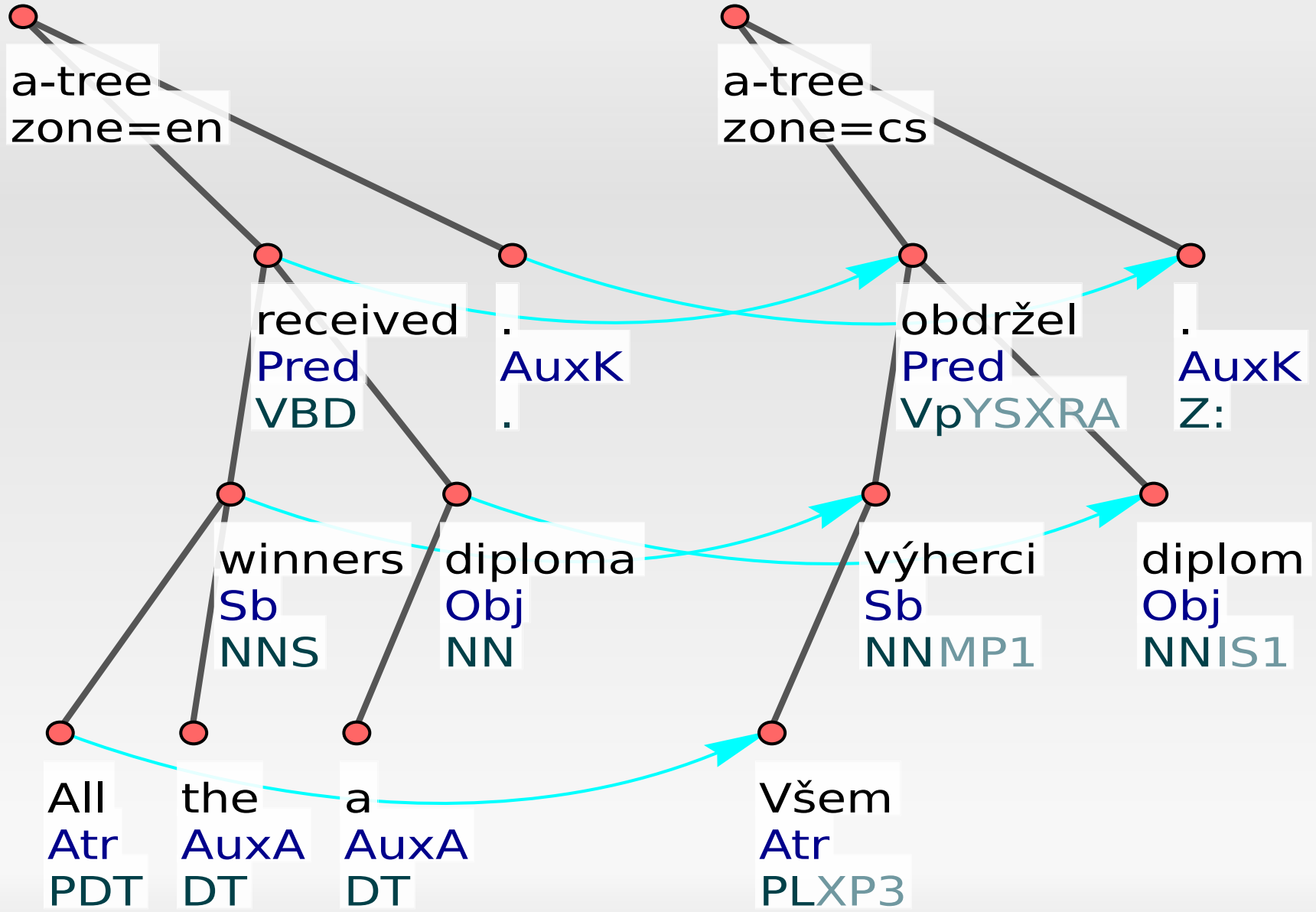
Podmět


a-tree
zone=en
a-tree
zone=cs
received
Pred
VBD
.
AuxK
.
obdržel
Pred
VpYSXRA
.
AuxK
Z:
winners
Sb
NNS
diploma
Obj
NN
výhercům
Obj
NNMP3
diplom
Obj
NNIS1
All
Atr
PDT
the
AuxA
DT
a
AuxA
DT
Všem
Atr
PLXP3

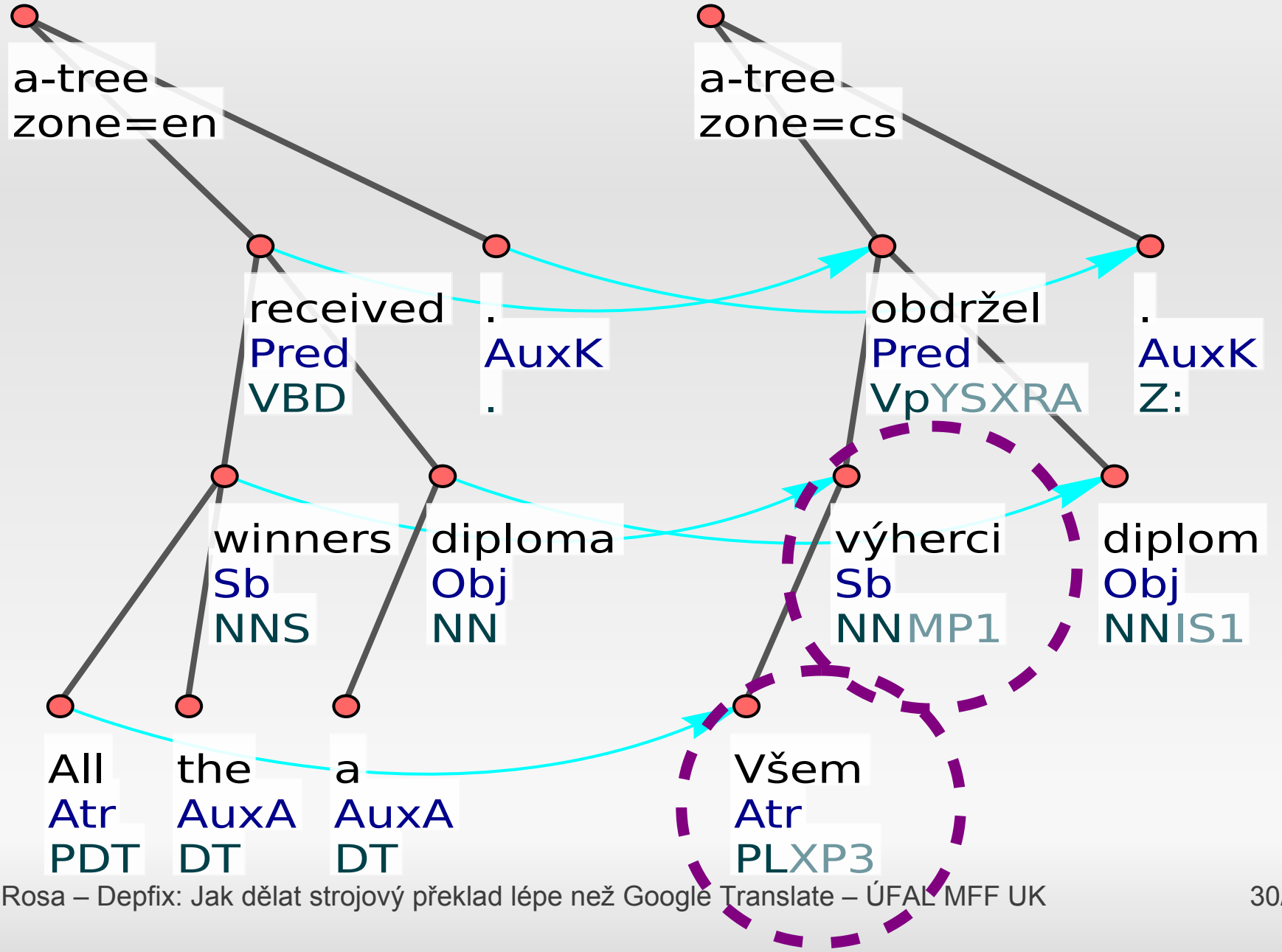
Podmět → 1. pád


a-tree
zone=en
received
Pred
VBD
.
AuxK
.
winners
Sb
NNS
diploma
Obj
NN
All
Atr
PDT
the
AuxA
DT
a
AuxA
DT
a-tree
zone=cs
obdržel
Pred
VpYSXRA
.
AuxK
Z:
výherci
Sb
NNMP1
diplom
Obj
NNIS1
Všem
Atr
PLXP3

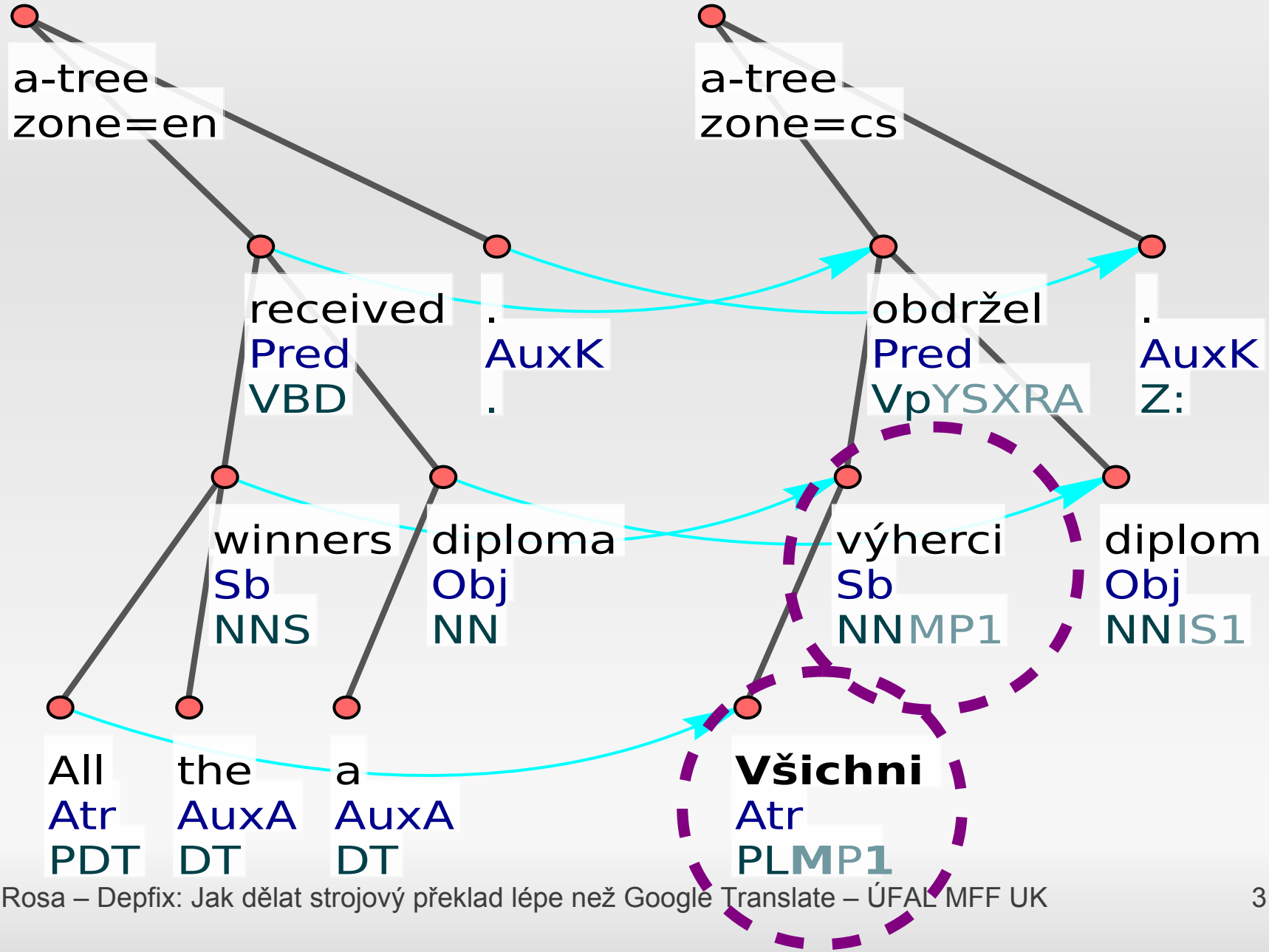
Všem výherci obdržel diplom.



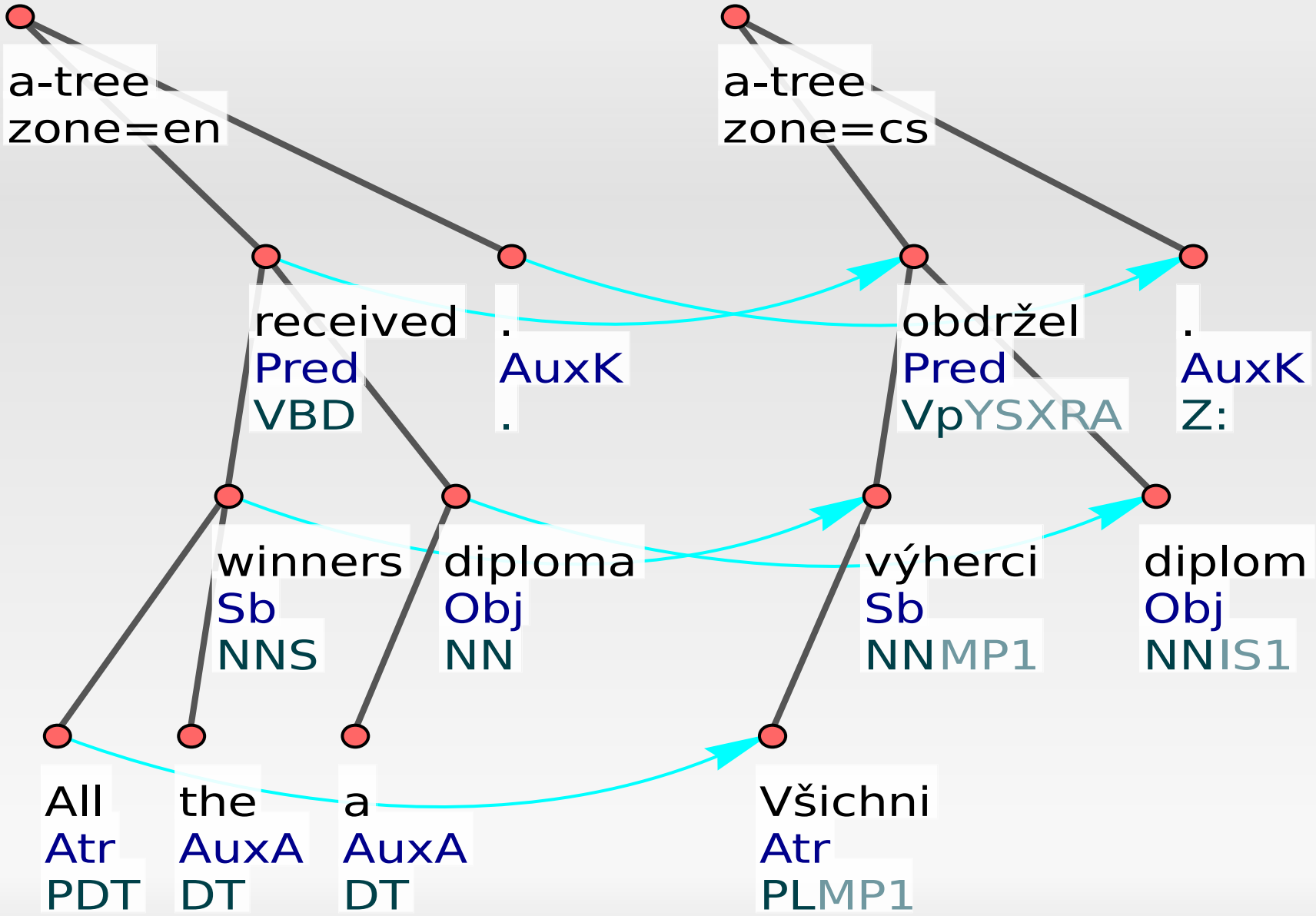
Shoda jména a přívlastku



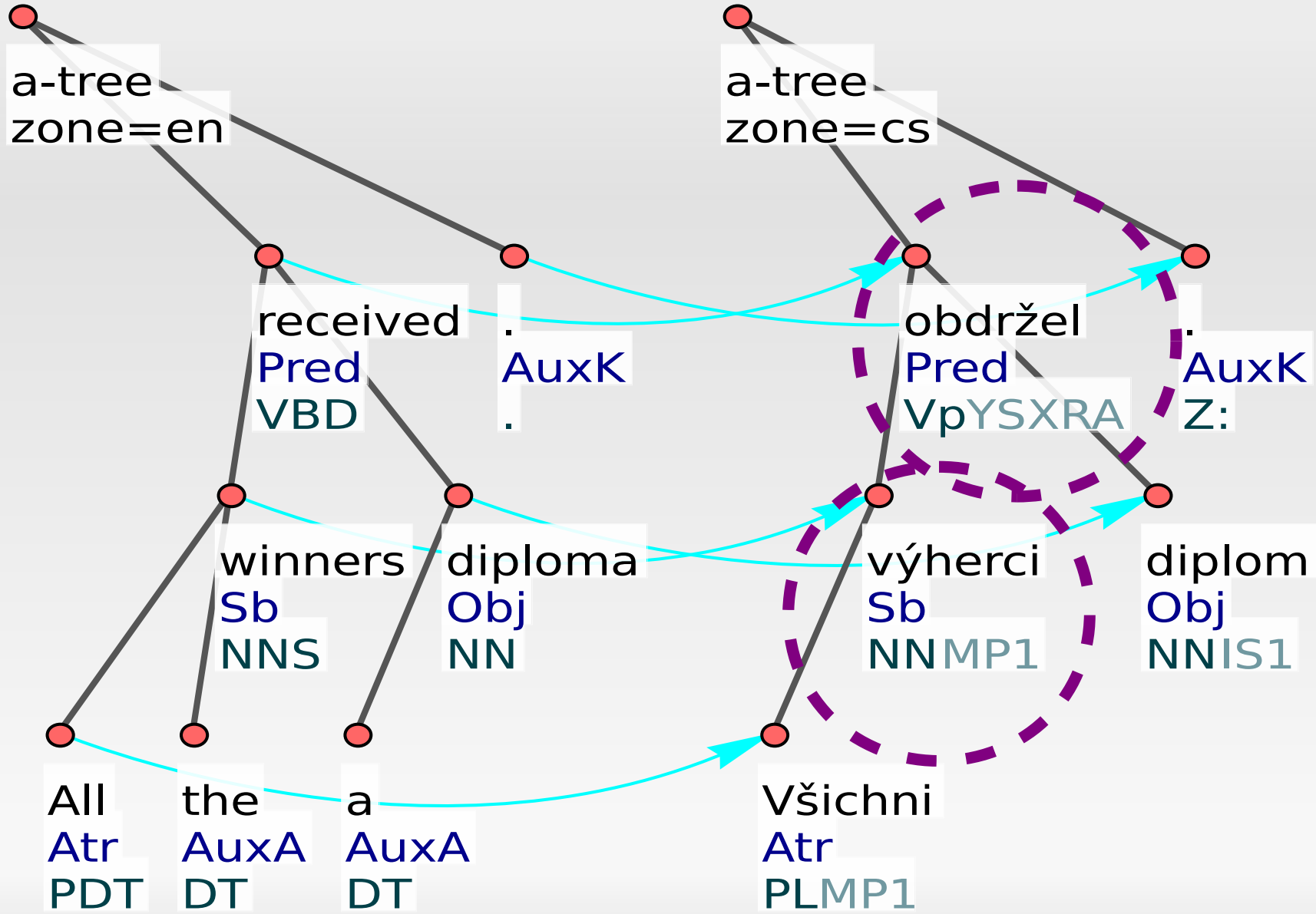
Shoda: rod, číslo, pád



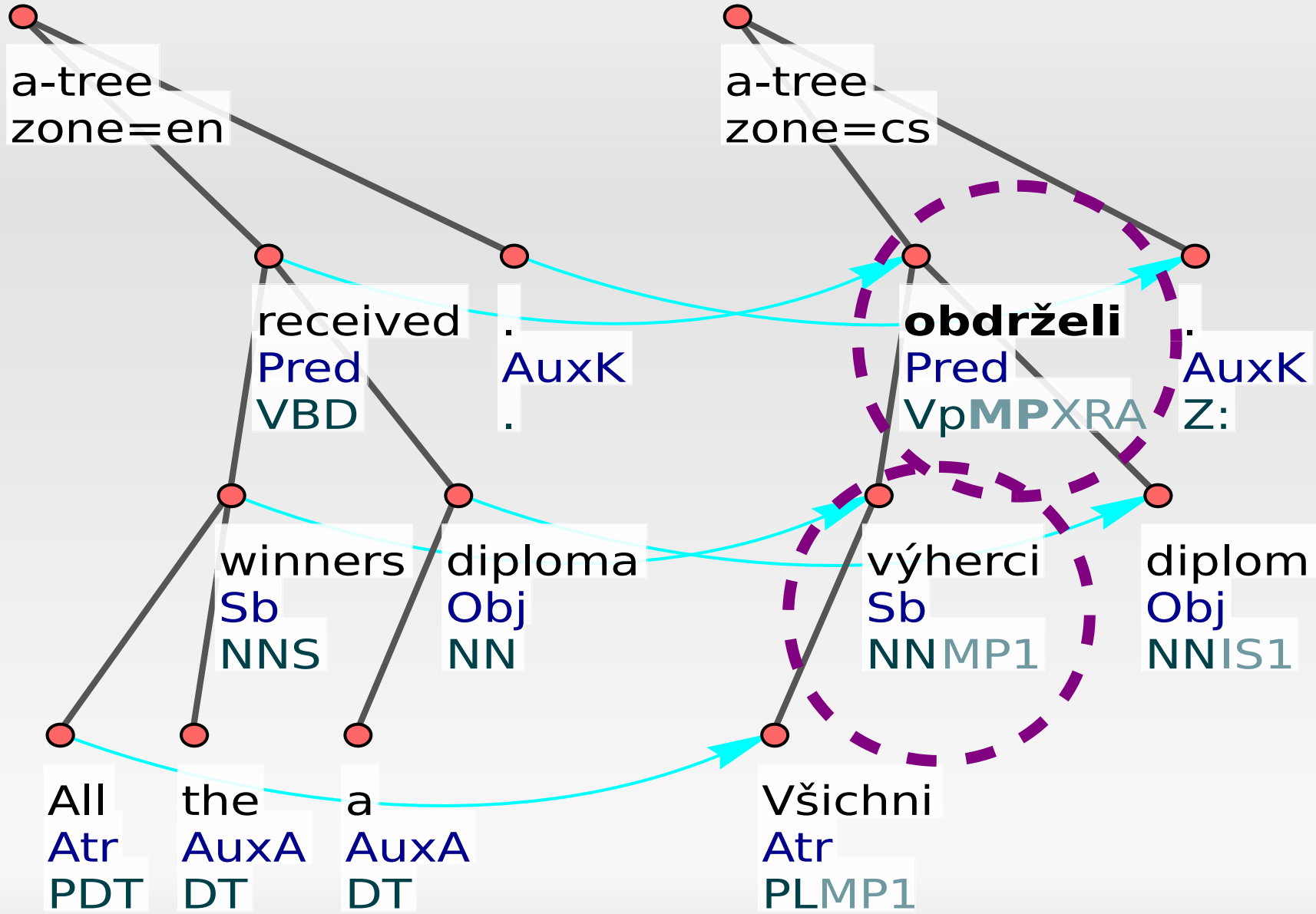
Všichni výherci obdržel diplom.



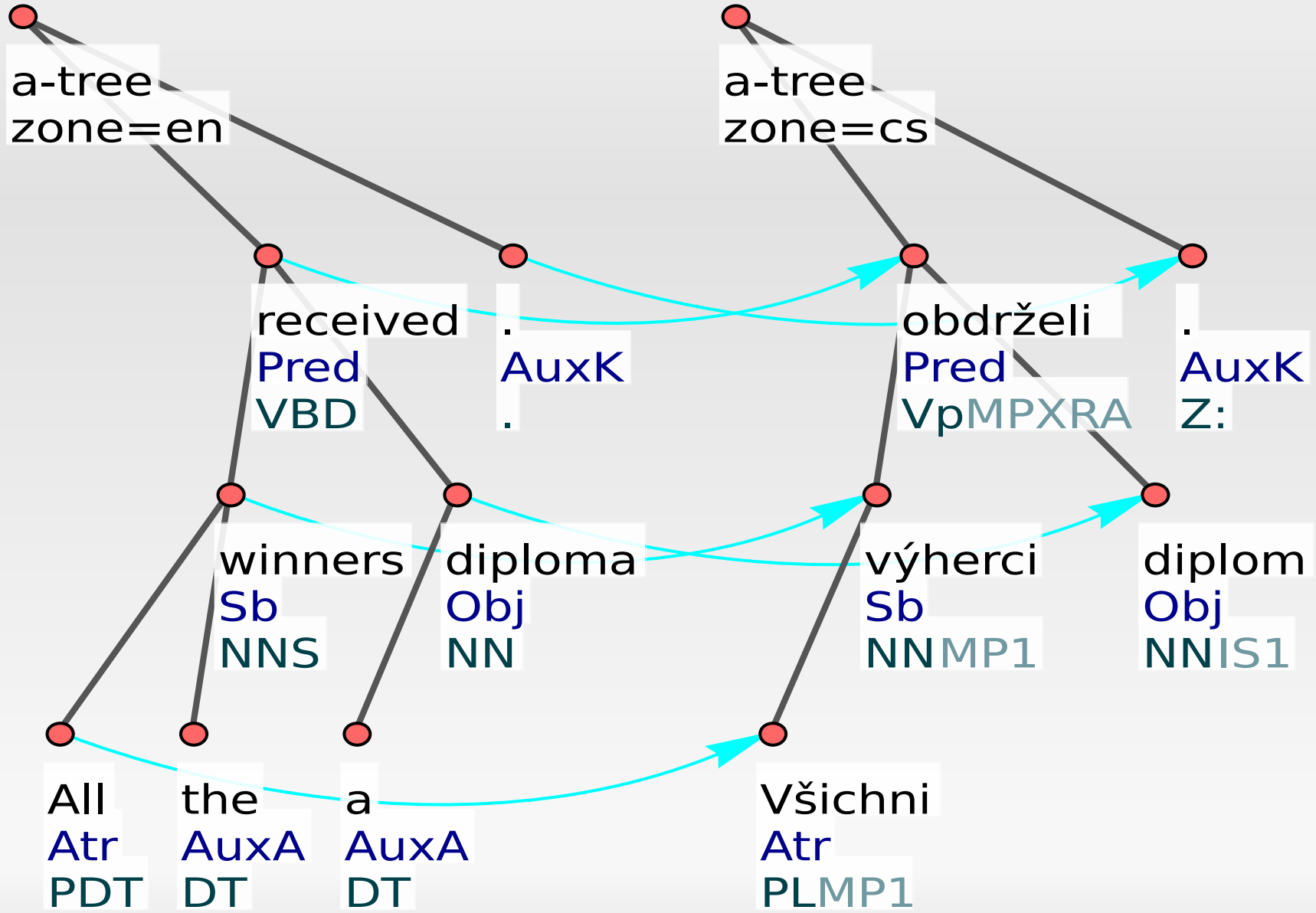
Shoda podmětu s přísudkem



Shoda: rod, číslo, osoba



Všichni výherci obdrželi diplom.



Jak Depfix opravuje chyby

- anglická věta:
  - *All the winners received a diploma.*
- překlad pomocí systému Moses:
  - *Všem výhercům obdržel diplom.*
- oprava pomocí systému Depfix:
  - *Všichni výherci obdrželi diplom.*

Oprava slovesné valence

- anglická věta:

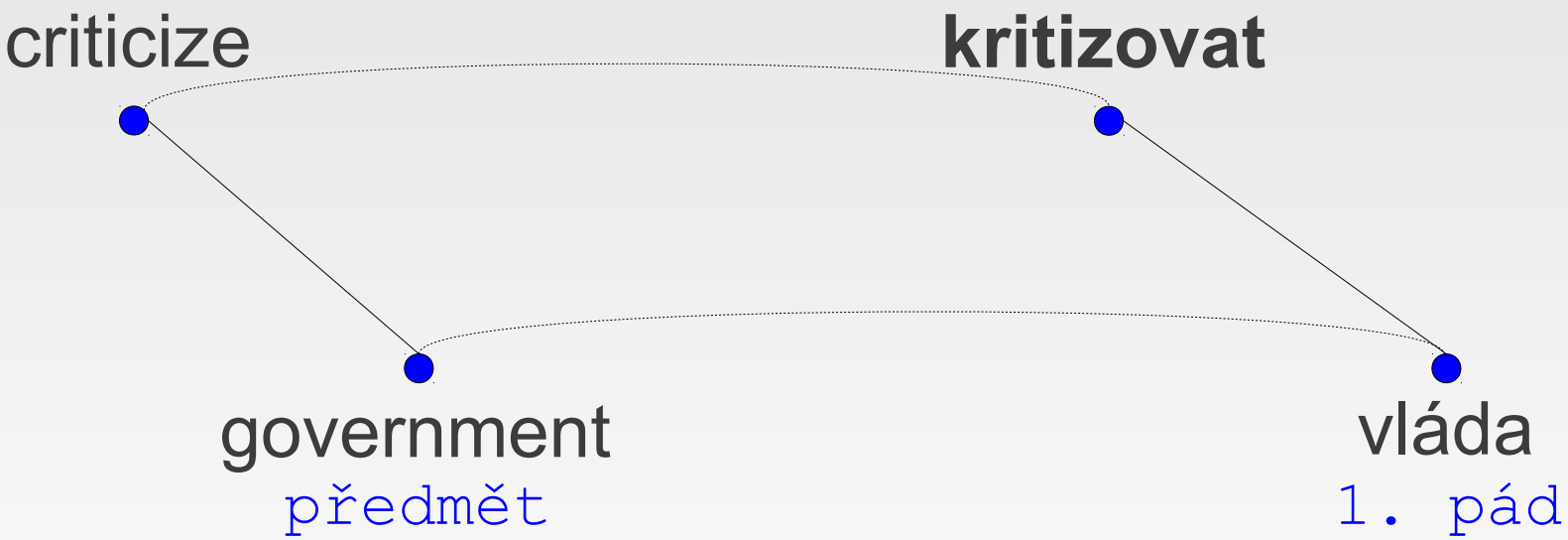
  EU criticizes not only the Greek government
- překlad pomocí Google Translate:

  EU kritizuje nejen ***řecká vláda***
- oprava pomocí systému Depfix:

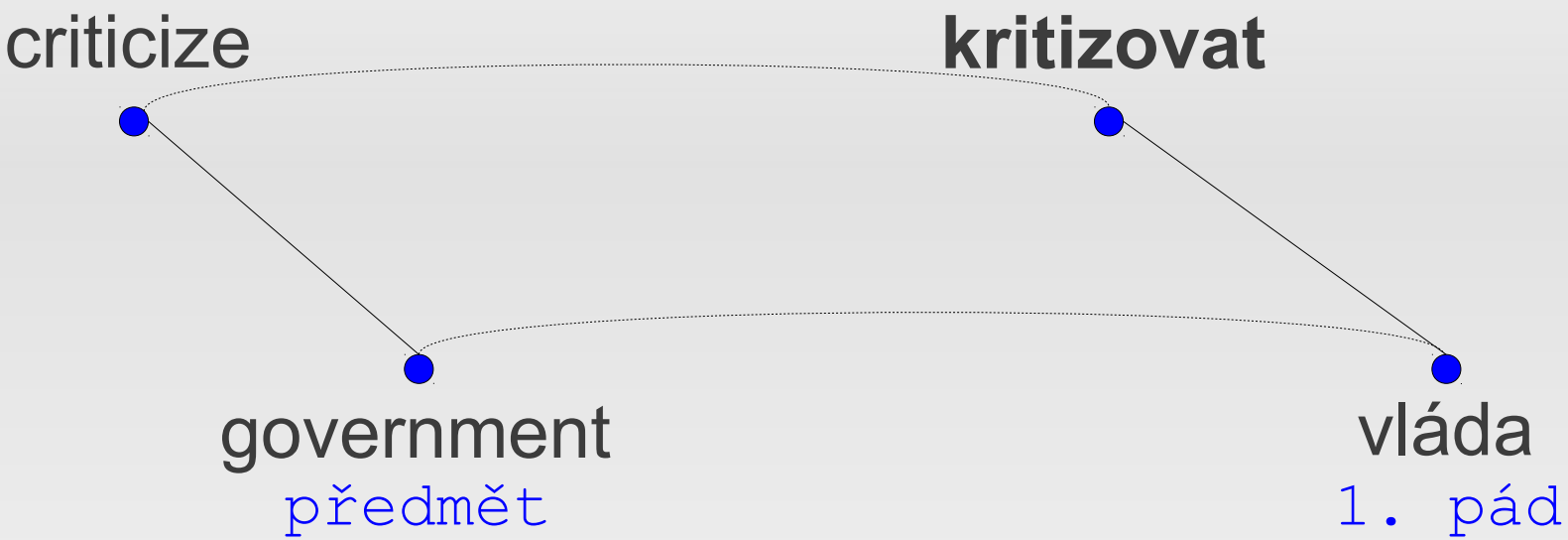
  EU kritizuje nejen ***řeckou vládu***

Automatický rozbor

- *EU criticizes not only the Greek government*
- *EU kritizuje nejen řecká vláda*

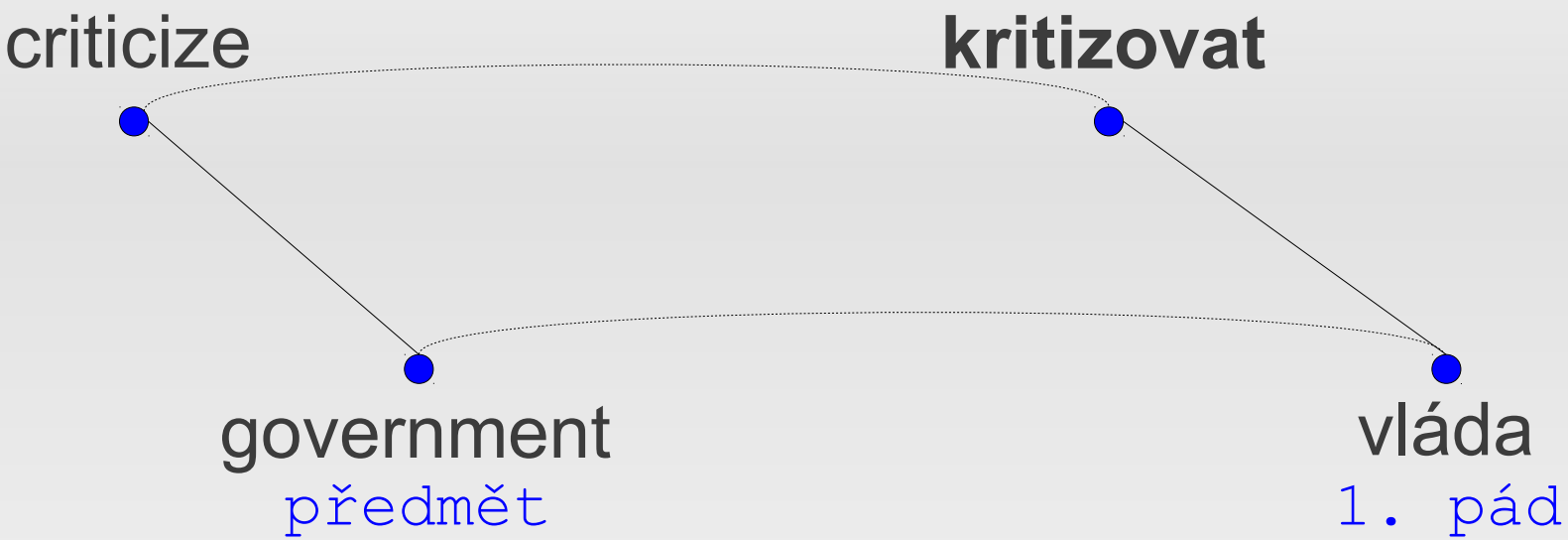


Pravděpodobnosti pádů

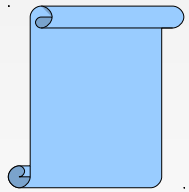


- sloveso **kritizovat**, pád pro **předmět**:
  - P(1. pád) = ?%
  - P(4. pád) = ?%
  - ...

Pravděpodobnosti pádů



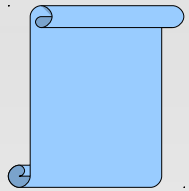
- sloveso **kritizovat**, pád pro **předmět**:
  - P(1. pád) = 3%
  - P(4. pád) = 80%



paralelní korpus

Pravděpodobnosti pádů

- sloveso **kritizovat**, pád pro **předmět**:
  - P(1. pád) = 3%
  - P(**4. pád**) = **80%**



paralelní korpus

- překlad pomocí Google Translate:

  *EU kritizuje nejen řecká **vláda***

- oprava pomocí systému Depfix:

  *EU kritizuje nejen řeckou **vládu***

Závěr

- frázový strojový překlad je jednoduchý a úspěšný
  - celkem se obejde bez lingvistiky
  - ale má problémy s gramatikou
- jeho chyby lze automaticky opravovat
  - na to se lingvistika hodí
  - automaticky určovat slovní druhy, větné členy...
- dva možné přístupy k opravě chyb
  - člověk ručně sestaví opravovací pravidla
  - systém se naučí pravidla sám (podle korpusu)

Přijďte k nám studovat!

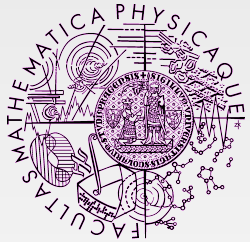
- bakalářské studium (3 roky)
  - Matematicko-fyzikální fakulta UK („Matfyz“)
  - obecná informatika, zaměření matematická lingvistika
  - http://www.mff.cuni.cz
- navazující magisterské studium (2 roky),
  pak případně doktorské studium (4 roky)
  - Ústav formální a aplikované lingvistiky (ÚFAL)
  - informatika, obor matematická lingvistika
  - http://ufal.mff.cuni.cz

Děkuji za pozornost

Rudolf Rosa
rur@nikde.eu

Depfix:

Jak dělat strojový překlad lépe než Google Translate



Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Pro tuto prezentaci a další informace navštivte

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