

4. PDT, SYNTAKTICKÁ ANOTACE V ČNK

Jakub Machura

Masarykova univerzita
Ústav českého jazyka
machura@phil.muni.cz

PDT PRAGUE DEPENDENCY TREEBANK

PDT PRAGUE DEPENDENCY TREEBANK

- 1. syntakticky anotovaný korpus češtiny

PDT PRAGUE DEPENDENCY TREEBANK

- 1. syntakticky anotovaný korpus češtiny
- vyvíjen od r. 1995 na ÚFAL UK

PDT PRAGUE DEPENDENCY TREEBANK

- 1. syntakticky anotovaný korpus češtiny
- vyvíjen od r. 1995 na ÚFAL UK
- koncepčně založen na přístupu FGP

PDT PRAGUE DEPENDENCY TREEBANK

- 1. syntakticky anotovaný korpus češtiny
- vyvíjen od r. 1995 na ÚFAL UK
- koncepčně založen na přístupu FGP
- obsahuje texty z ČNK

PDT PRAGUE DEPENDENCY TREEBANK

- 1. syntakticky anotovaný korpus češtiny
- vyvíjen od r. 1995 na ÚFAL UK
- koncepčně založen na přístupu FGP
- obsahuje texty z ČNK
- anotace na 3 úrovních popisu

PDT PRAGUE DEPENDENCY TREEBANK

- 1. syntakticky anotovaný korpus češtiny
- vyvíjen od r. 1995 na ÚFAL UK
- koncepčně založen na přístupu FGP
- obsahuje texty z ČNK
- anotace na 3 úrovních popisu
- 3 165 dokumentů, 49 431 vět

PDT PRAGUE DEPENDENCY TREEBANK

- morfologická rov.

PDT PRAGUE DEPENDENCY TREEBANK

- morfologická rov.
 - 2 mil. slovních výskytů

PDT PRAGUE DEPENDENCY TREEBANK

- morfologická rov.
 - 2 mil. slovních výskytů
- analytická rov.

PDT PRAGUE DEPENDENCY TREEBANK

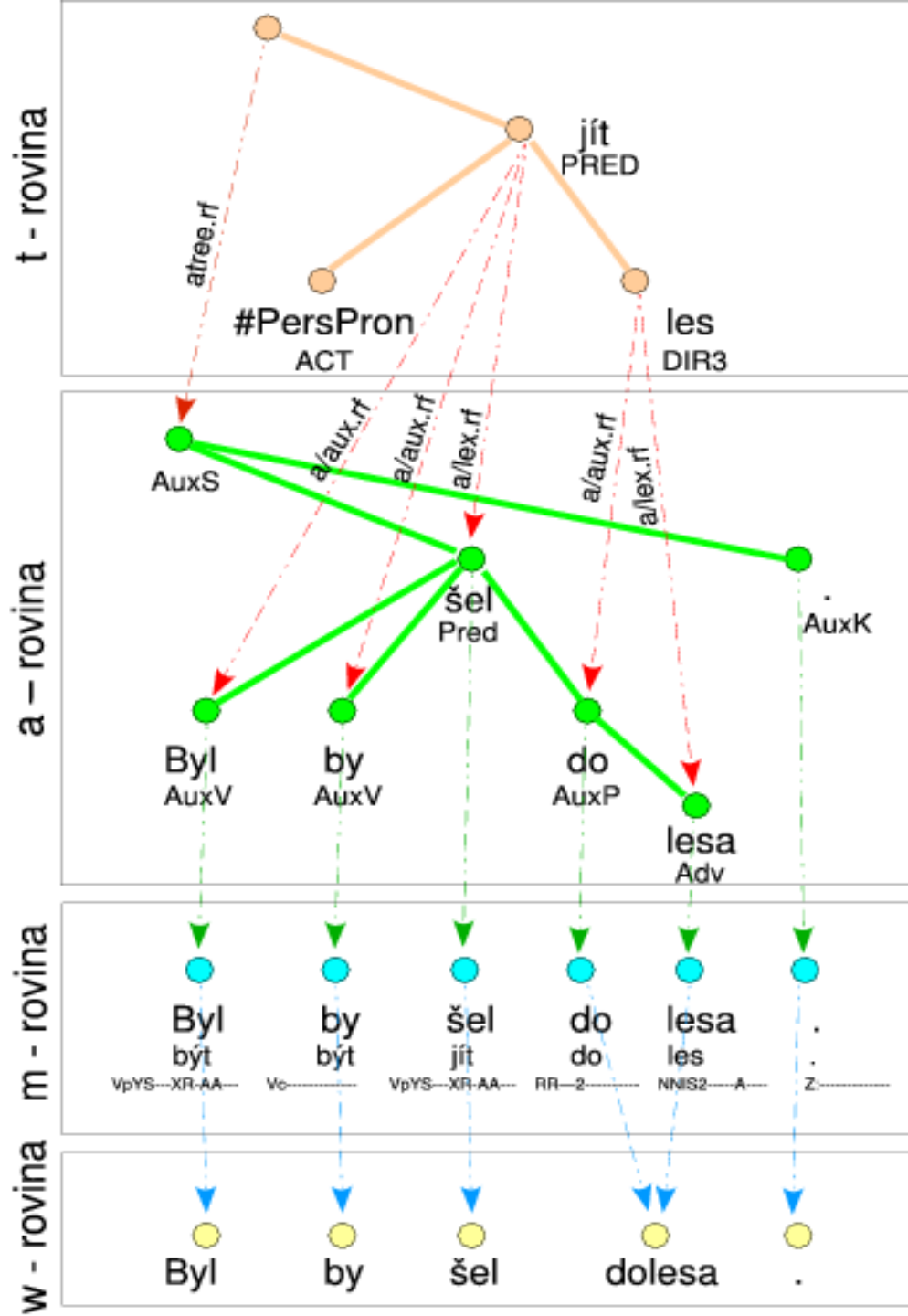
- morfologická rov.
 - 2 mil. slovních výskytů
- analytická rov.
 - 1,5 mil. slovních tvarů

PDT PRAGUE DEPENDENCY TREEBANK

- morfologická rov.
 - 2 mil. slovních výskytů
- analytická rov.
 - 1,5 mil. slovních tvarů
- tektogramatická rov.

PDT PRAGUE DEPENDENCY TREEBANK

- morfologická rov.
 - 2 mil. slovních výskytů
- analytická rov.
 - 1,5 mil. slovních tvarů
- tektogramatická rov.
 - 833 tis. uzlů



PDT PRAGUE DEPENDENCY TREEBANK

- větné členy (7)

PDT PRAGUE DEPENDENCY TREEBANK

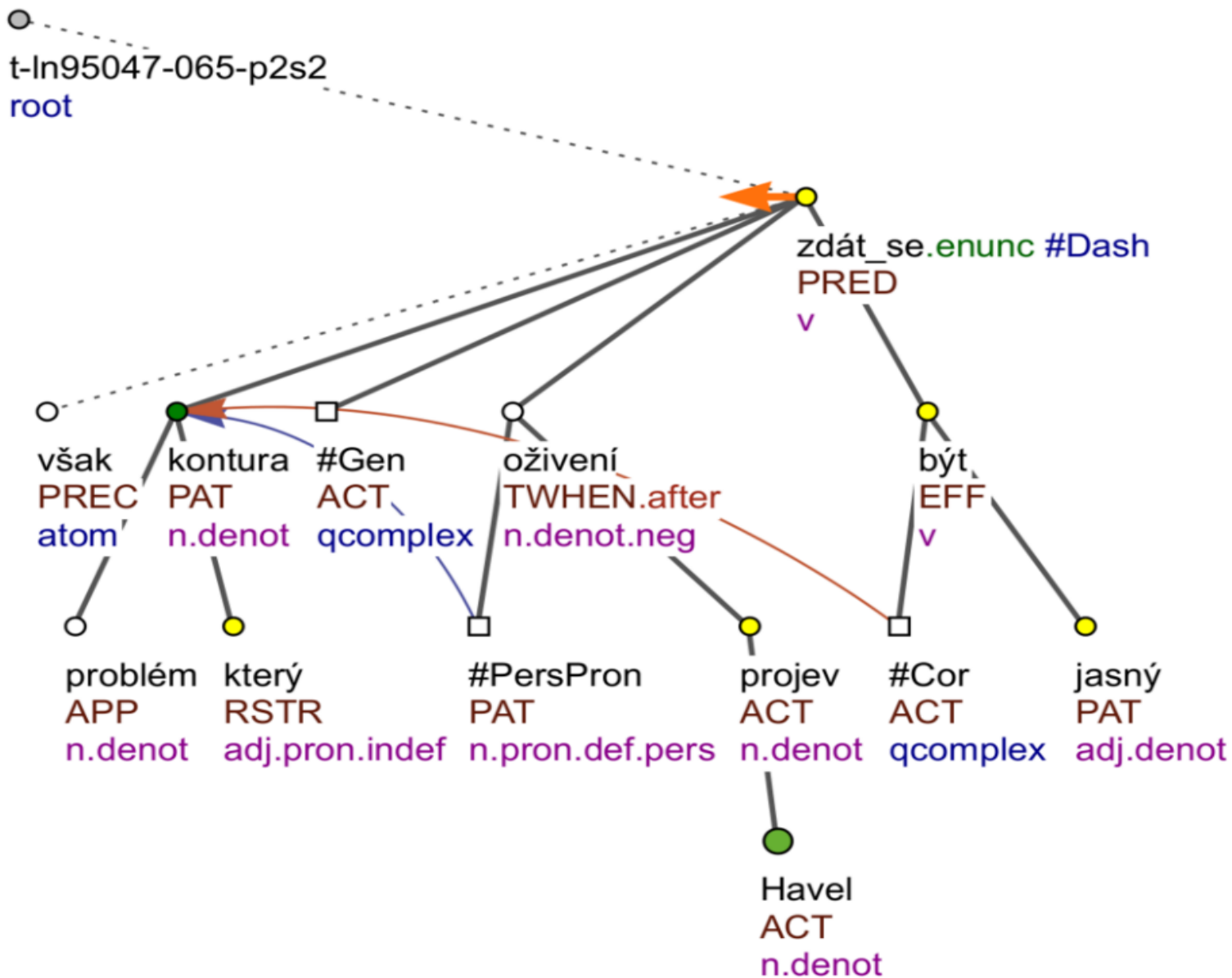
- větné členy (7)
- funktor (44), subfunktory

PDT PRAGUE DEPENDENCY TREEBANK

- větné členy (7)
- funktor (44), subfunktory
- aktanty a volná doplnění

PDT PRAGUE DEPENDENCY TREEBANK

- větné členy (7)
- funktor (44), subfunktory
- aktanty a volná doplnění
- koordinace, apozice, parenteze



Některé kontury problému se však po oživení Havlovým projevem zdají být jasnější.

PDT PRAGUE DEPENDENCY TREEBANK

- vhodná materiálová základna pro lingvistické studie

PDT PRAGUE DEPENDENCY TREEBANK

- vhodná materiálová základna pro lingvistické studie
- NLP – použití statistických metod a strojového učení

PDT PRAGUE DEPENDENCY TREEBANK

- vhodná materiálová základna pro lingvistické studie
- NLP – použití statistických metod a strojového učení
- PDiT (*Prague Discourse Treebank*)

PDT PRAGUE DEPENDENCY TREEBANK

- vhodná materiálová základna pro lingvistické studie
- NLP – použití statistických metod a strojového učení
- PDiT (*Prague Discourse Treebank*)
- PCEDT (*Prague Czech-English Dependency Treebank*)

PDT PRAGUE DEPENDENCY TREEBANK

- vhodná materiálová základna pro lingvistické studie
- NLP – použití statistických metod a strojového učení
- PDiT (*Prague Discourse Treebank*)
- PCEDT (*Prague Czech-English Dependency Treebank*)
- PDTSC (*Prague Dependency Treebank of Spoken Language*)

SYNTAKTICKÁ ANOTACE V ČNK

SYNTAKTICKÁ ANOTACE V ČNK

- SYN2015

SYNTAKTICKÁ ANOTACE V ČNK

- SYN2015
- vychází ze zásad na tzv. analytické rovině v PDT
(1,5 mil. kvalitně označkových dat)

SYNTAKTICKÁ ANOTACE V ČNK

- SYN2015
- vychází ze zásad na tzv. analytické rovině v PDT (1,5 mil. kvalitně označkových dat)
- v korpusech ČNK automaticky (TurboParser)

SYNTAKTICKÁ ANOTACE V ČNK

- SYN2015
- vychází ze zásad na tzv. analytické rovině v PDT (1,5 mil. kvalitně označovaných dat)
- v korpusech ČNK automaticky (TurboParser)
- 1/6 tokenů označena chybně

SYNTAKTICKÁ ANOTACE V ČNK

- SYN2015
- vychází ze zásad na tzv. analytické rovině v PDT (1,5 mil. kvalitně označovaných dat)
- v korpusech ČNK automaticky (TurboParser)
- 1/6 tokenů označena chybně
- přiřazení rodiče: 88,48 %

SYNTAKTICKÁ ANOTACE V ČNK

- SYN2015
- vychází ze zásad na tzv. analytické rovině v PDT (1,5 mil. kvalitně označovaných dat)
- v korpusech ČNK automaticky (TurboParser)
- 1/6 tokenů označena chybně
- přiřazení rodiče: 88,48 %
- přiřazení rodiče a synt. funkce: 82,46 %

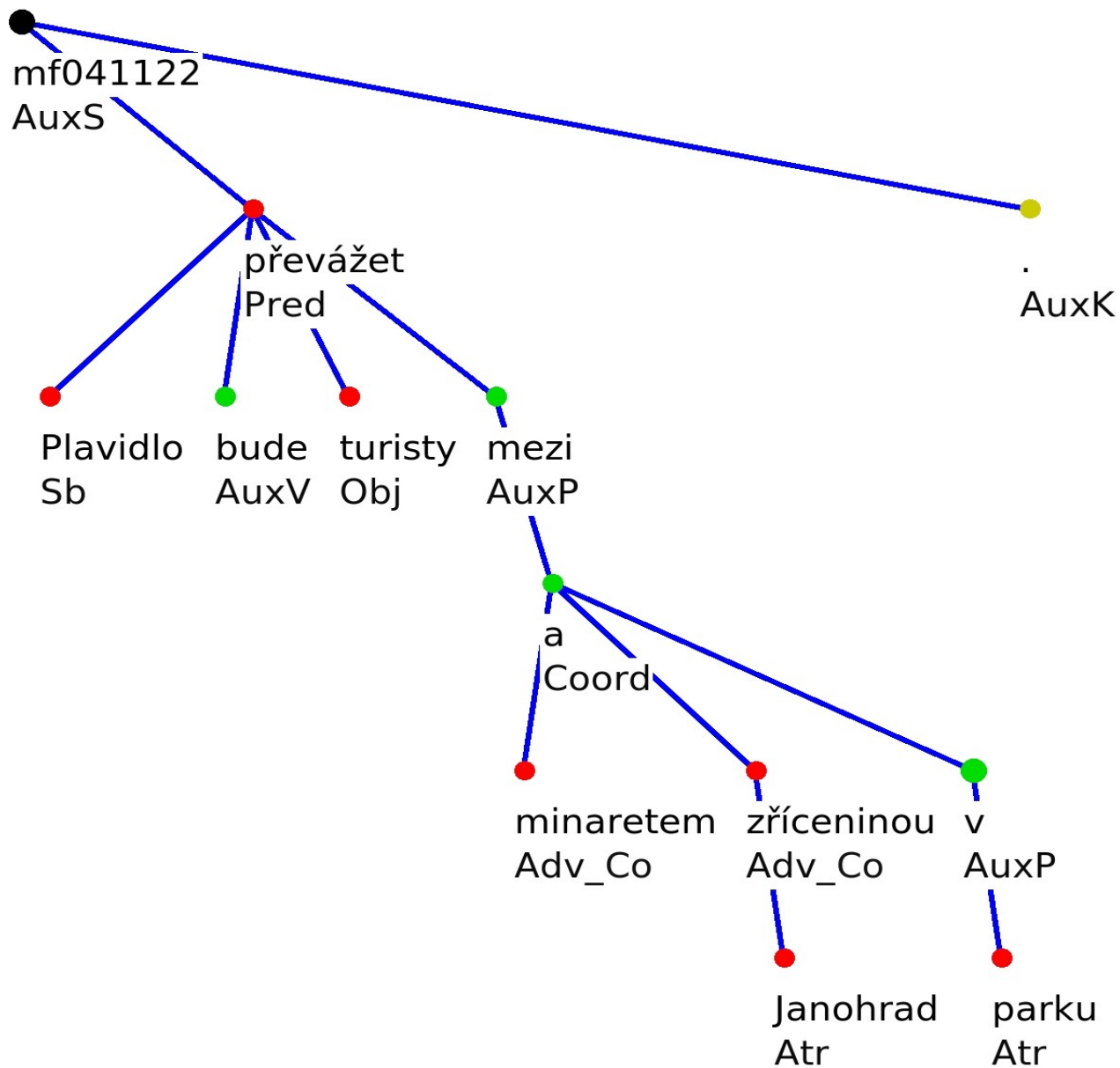
SYNTAKTICKÁ ANOTACE V ČNK

- každému tokenu přiřazen „rodič“
- každému tokenu přiřazena syntaktická značka (funkce)

Afun	Definice	Poznámka
<u>Pred</u>	Predikát (přísudek)	pouze v hlavní větě
<u>Sb</u>	Subjekt (podmět)	
<u>Obj</u>	Objekt (předmět)	včetně určení činitele děje u opisného pasiva; vč. infinitivu u modálních a fázových sloves aj.
<u>Adv</u>	Adverbiale (přísluvečné určení)	bez dalšího rozlišení, vč. předložkových pádů u slovesa <i>být</i> , některých volných dativů aj.
<u>Atv</u>	Doplněk	jen tzv. určující; technicky zavěšen na neslovesném členu
<u>AtvV</u>	Doplněk	jen tzv. určující; visící na slovese (ve větě není vyjádřen druhý řídicí člen)
<u>Atr</u>	Atribut (přívlastek)	včetně spojení typu <i>Jan Novák</i> , genitivu po číslovkách aj.
<u>Pnom</u>	Jmenná část verbonom. přísudku	se sponou <i>být, bývat</i> ; mimo předložkové pády

Afun	Definice	Poznámka
<u>AuxV</u>	Pomocné sloveso <i>být</i>	(Auxiliary Verb)
<u>Coord</u>	Koordinační uzel	uzel, který reprezentuje souřadící spojení: souřadící spojka nebo interpunkce
<u>Apos</u>	Aposice (hlavní uzel)	spojovací slovo nebo interpunkční znaménko
<u>AuxT</u>	Zvratné se u refl. tantum	neoddělitelné se – reflexivní tantum
<u>AuxR</u>	Zvratné se u pasiva	zvratné se, které není Obj ani AuxT (tvoří pasivum reflexivní)
<u>AuxP</u>	Předložka	popř. všechny části složené předložky
<u>AuxC</u>	Spojka podřadící	

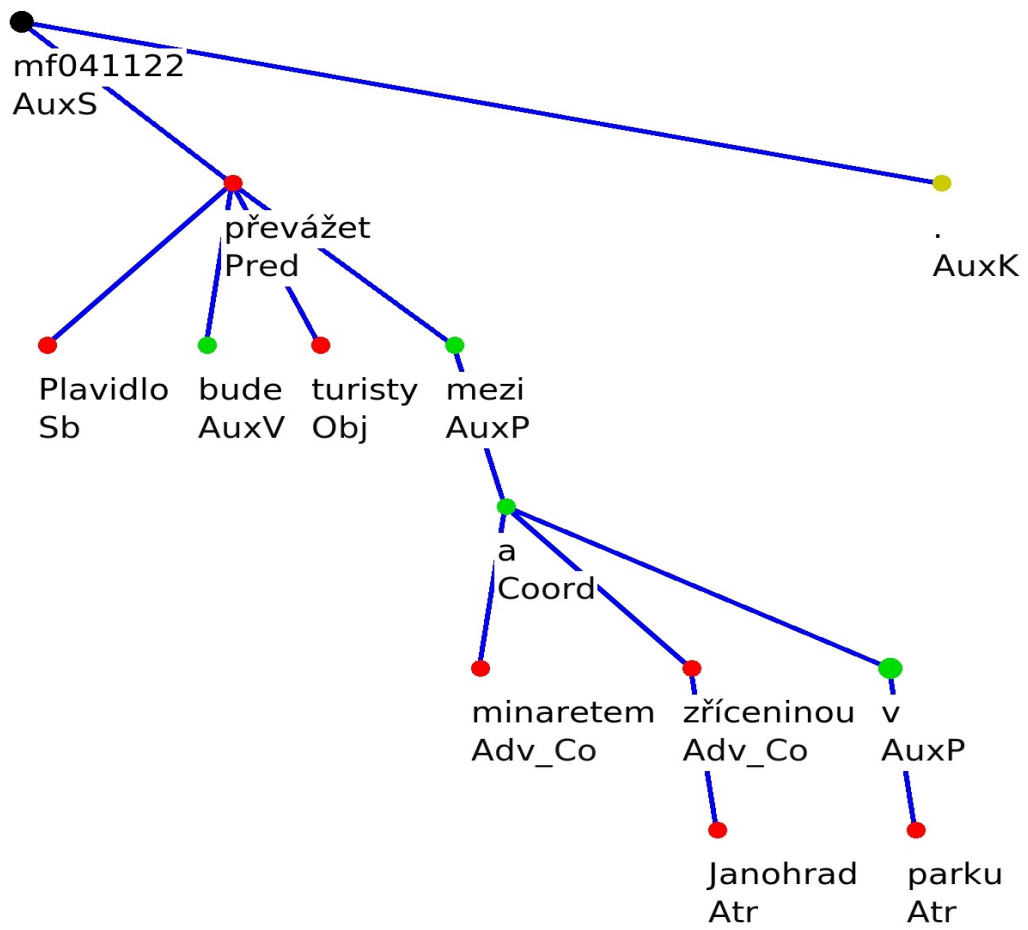
Afun	Definice	Poznámka
<u>AuxO</u>	Nadbytečný element	odkazovací, emotivní element ap., např. etické <i>ti</i>
<u>AuxZ</u>	Zdůrazňovací slovo	
<u>AuxX</u>	Čárka	ne však nositel koordinace nebo koncový symbol
<u>AuxG</u>	Jiné grafické symboly, které neukončují větu	
<u>AuxY</u>	Příslovce a částice, které nelze zařadit jinam	
<u>AuxK</u>	Koncová interpunkce věty	
<u>ExD</u>	Součást aktuální elipsy	náhradní funkce pro technické hrany vedoucí místo od elidovaného členu k „pseudořídícímu“ slovu nebo pro hlavní člen věty bez predikátu (Ex-Dependent); také ve srovnávacích spojeních typu <i>zdravý jako ryba</i>



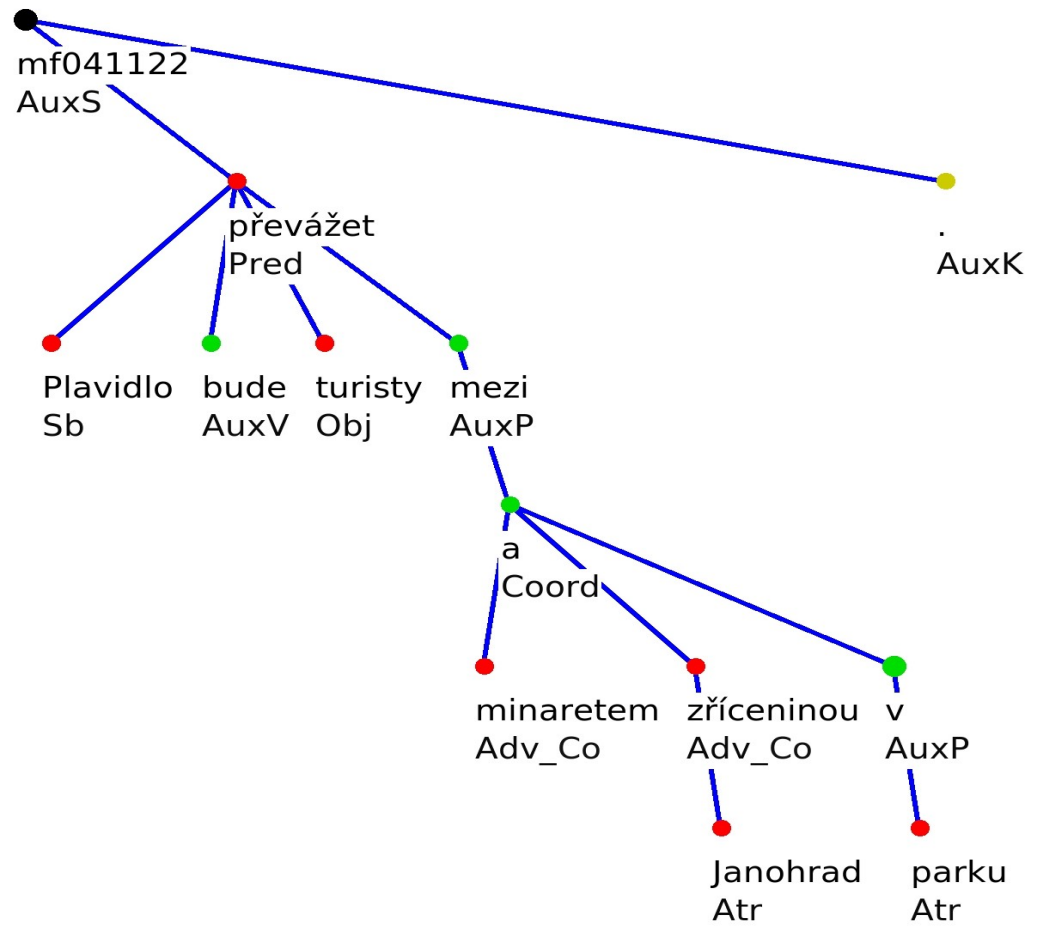
Plavidlo bude převážet turisty mezi minaretem a zříceninou Janohrad v parku.

SYNTAKTICKÁ ANOTACE V ČNK

- lze vyhledávat slova a slovní spojení podle syntaktických parametrů
- parent = číselný odkaz na pozici řídicího členu
- afun = syntaktická funkce

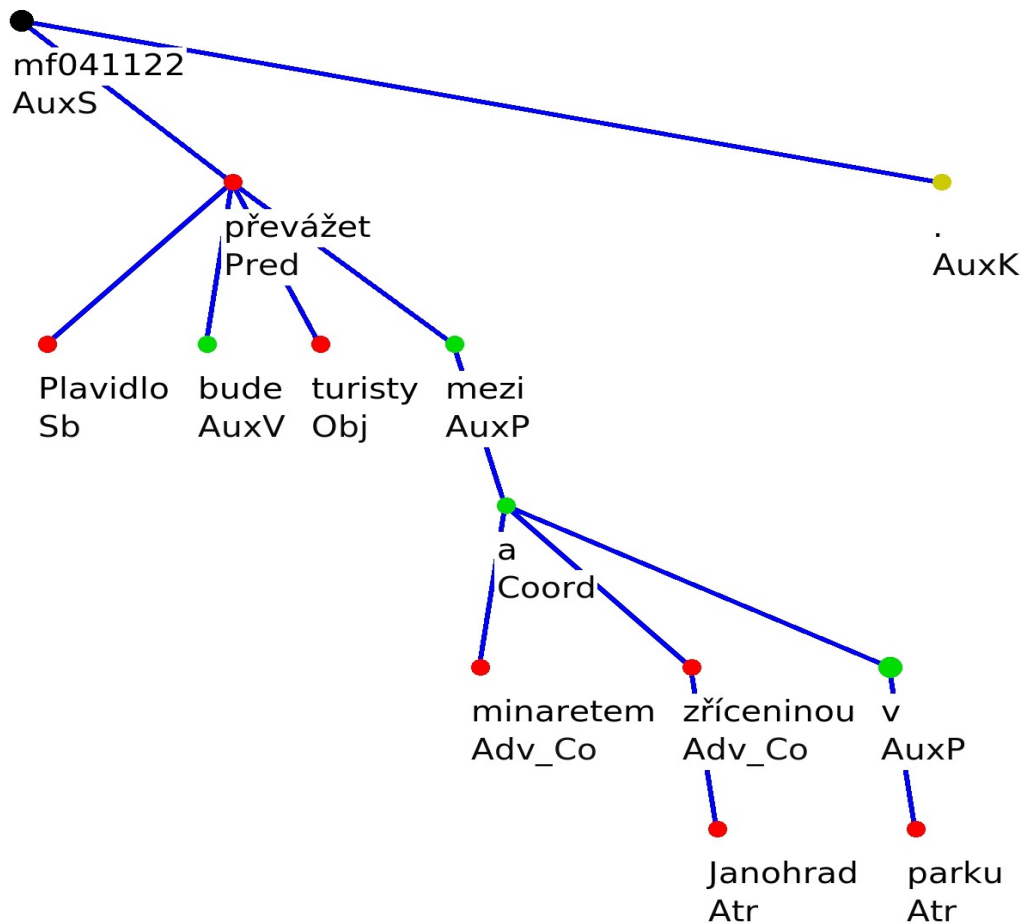


afun="Adv_Co"



afun="Adv_Co"

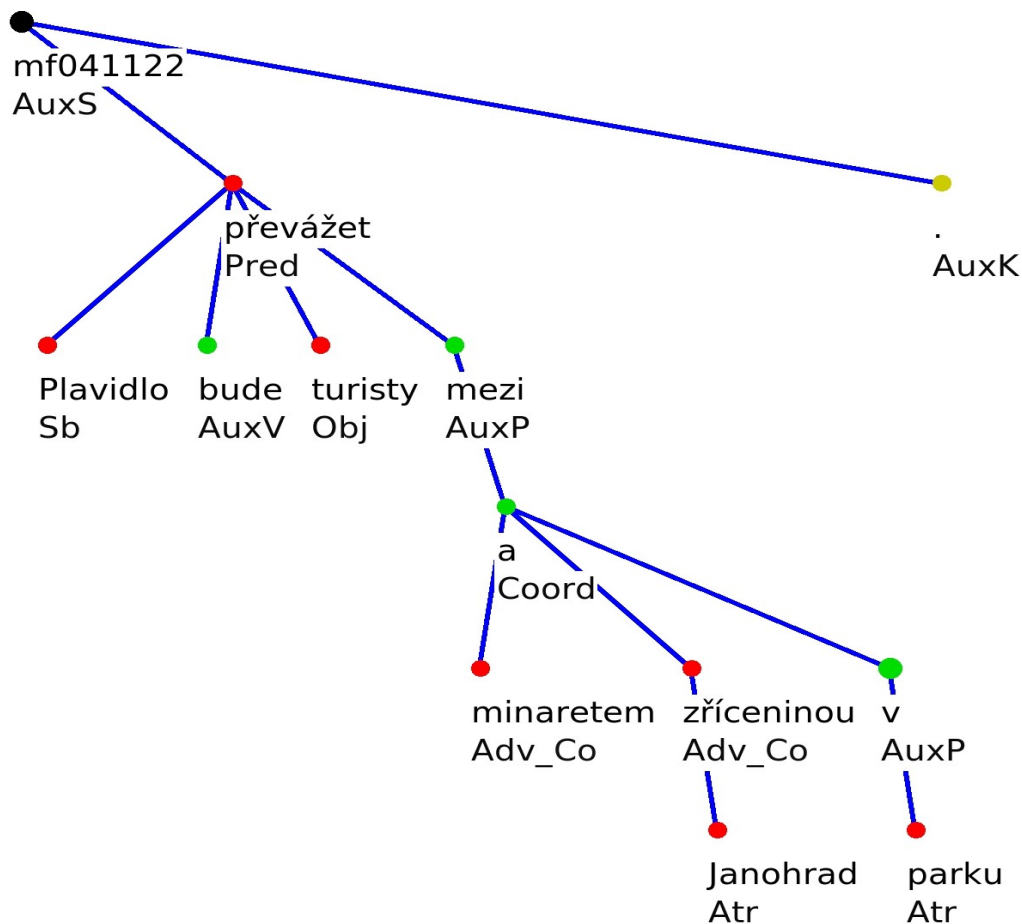
parent="-1",
p_tag="J^-----",
p_lemma="a",
p_afun="Coord";



afun="Adv_Co"

parent="-1",
p_tag="J^-----",
p_lemma="a",
p_afun="Coord";

eparent="-5,
"ep_tag="Vf-----",
ep_lemma="převážet",
ep_afun="Pred";



SYNTAKTICKÁ ANOTACE V ČNK

- V korpusu pak lze podle těchto atributů vyhledávat, např. lze vyhledat všechna substantiva v akuzativu se syntaktickou funkcí Obj závislá na slovese *převážet*:

SYNTAKTICKÁ ANOTACE V ČNK

- V korpusu pak lze podle těchto atributů vyhledávat, např. lze vyhledat všechna substantiva v akuzativu se syntaktickou funkcí Obj závislá na slovese *převážet*:

```
[afun="Obj" & tag="NN..4.*" & p_lemma="převážet"]
```

SYNTAKTICKÁ ANOTACE V ČNK

- Nebo lze vyhledat všechna slova (syntaktická substantiva) v sedmém pádě s předložkou *mezi* závislá na slovese v infinitivu:

SYNTAKTICKÁ ANOTACE V ČNK

- Nebo lze vyhledat všechna slova (syntaktická substantiva) v sedmém pádě s předložkou *mezi* závislá na slovese v infinitivu:

```
[prep="mezi" & case="7" & ep_tag="Vf.*"]
```

SYNTAKTICKÁ ANOTACE V ČNK

Vyhledávat lze ve směru od závislého tokenu k řídícímu, ne naopak!

SYNTAKTICKÁ ANOTACE V ČNK

- z morfologické anotace lze získat mnoho poznatků o syntaxi
- zakódovaná syntax – např. při desambiguaci

SYNTAKTICKÁ ANOTACE V ČNK

Případová studie:

Reflexivní slovesa v češtině s akuzativní valencí

Která reflexivní (zvratná) slovesa v češtině mohou mít předmět v prostém akuzativu?

SYNTAKTICKÁ ANOTACE V ČNK

Případová studie:

Reflexivní slovesa v češtině s akuzativní valencí

Která reflexivní (zvratná) slovesa v češtině mohou mít předmět v prostém akuzativu?

```
[tag="V.*"] [tag="V[Bc].*" & lemma="být"] ?
```

```
[pos="P.*" & lemma="se|si"]
```

```
[tag="[APC]...4.*"] * [tag="N...4.*"]
```

SYNTAKTICKÁ ANOTACE V ČNK

Shoda atributů v CQL

1. Pozice označíme číslem a dvojtečkou před hranatou závorkou
2. 1:[pos="A"] 2:[pos="A"]
3. Shodu atributů testujeme za amp. &
4. & 1.lemma=2.lemma

SYNTAKTICKÁ ANOTACE V ČNK

Shoda atributů v CQL

1. Pozice označíme číslem a dvojtečkou před hranatou závorkou
 2. 1:[pos="A"] 2:[pos="A"]
 3. Shodu atributů testujeme za amp. &
 4. & 1.lemma=2.lemma
- Najdi sekvenci substantivum, slovní tvar a a substantivum, přičemž lemmata obou substantiv jsou shodná.

SYNTAKTICKÁ ANOTACE V ČNK

Shoda atributů v CQL

1. Pozice označíme číslem a dvojtečkou před hranatou závorkou
2. 1:[pos="A"] 2:[pos="A"]
3. Shodu atributů testujeme za amp. &
4. & 1.lemma=2.lemma

- Najdi sekvenci substantivum, slovní tvar *a* a substantivum, přičemž lemmata obou substantiv jsou shodná.

1:[pos="N"] [word="a"] 2:[pos="N"] & 1.lemma = 2.lemma

SYNTAKTICKÁ ANOTACE V ČNK

NEshoda atributů v CQL

1. Pozice označíme číslem a dvojtečkou před hranatou závorkou
2. 1:[word="je"] 2:[word="je"]
3. NEshodu atributů testujeme za amp. &
4. & 1.pos!=2.pos

SYNTAKTICKÁ ANOTACE V ČNK

NEshoda atributů v CQL

1. Pozice označíme číslem a dvojtečkou před hranatou závorkou
2. 1:[word="je"] 2:[word="je"]
3. NEshodu atributů testujeme za amp. &
4. & 1.pos!=2.pos

Najdi dvě substantiva oddělená slovem *a*, která se **neshodují** v pádě

SYNTAKTICKÁ ANOTACE V ČNK

NEshoda atributů v CQL

1. Pozice označíme číslem a dvojtečkou před hranatou závorkou
2. 1:[word="je"] 2:[word="je"]
3. NEshodu atributů testujeme za amp. &
4. & 1.pos!=2.pos

Najdi dvě substantiva oddělená slovem *a*, která se **neshodují** v pádě

```
1:[pos="N"] [word="a"] 2:[pos="N"] & 1.case != 2.case
```


(NE)SHODA ATRIBUTŮ V CQL

3. Najděte dvě po sobě stojící slova se stejnou formou, ale odlišným lemmatem.
4. Najděte po sobě jdoucí předložku, případně další předložku, dále substantivum, případně adjektivum a druhé substantivum, přičemž 1. subst. má jiný pád než předložka, 2. subst. stejný.



(NE)SHODA ATRIBUTŮ V CQL

3. Najděte dvě po sobě stojící slova se stejnou formou, ale odlišným lemmatem.

1:[] 2:[] & 1.word = 2.word & 1.lemma != 2.lemma

4. Najděte po sobě jdoucí předložku, případně další předložku, dále substantivum, případně adjektivum a druhé substantivum, přičemž 1. subst. má jiný pád než předložka, 2. subst. stejný.



ZOBRAZENÍ ZÁVISLOSTNÍCH STRUKTUR

Výskytů: 106 | i.p.m. ⁰: 0,88 (vztaženo k celému "syn2015") | ARF ⁰: 65,58 | Výsledek je promíchán

Výběr řádků: základní | Atributy:

<input type="checkbox"/>	Kloktat dehet	kde jsme s Martinem rvali mříž , padala omítka .	Mariánský zpěváci tahali za šňůru .
<input type="checkbox"/>	Svět motorů	jinak tak čiperné , ani nechtělo sbírat ze zatáček .	Mdlý zvuk nestojí za řeč .
<input type="checkbox"/>	Valdštejn a Lukrecie	dodá hlasem skoro výhrůžným : „ Zůstaneš tady ! “	Teplá tma padá na Valdštejna .
<input type="checkbox"/>	Hospodářské noviny	šetří a propouští NA POZADÍ SLABŠÍCH POLOLETNÍCH VÝSLEDKŮ SE ČESKÁ	ENERGETICKÁ SKUPINA PŘIPRAVUJE NA MASIVNÍ ŠKI
<input type="checkbox"/>	Hospodářské noviny	rámci sporu izraelského státu s dědičkami o vlastnictví pozůstalosti .	Státní zaměstnanci přijdou o tisíce .
<input type="checkbox"/>	Klikněte pro zobrazení syntaktického stromu	rbonská ; sedimentace přetrvala lokálně až do středního triasu .	Největší podíl připadá na permské sedimenty .
<input type="checkbox"/>	Krajské noviny	v jakém volebním obvodu kandiduje , “ prohlásil Antl .	Zajímavá osobnost kandiduje za ústecký obvod .
<input type="checkbox"/>	Čistý	velmi ulevilo . Obě okénka drožky jsou úplně stažená .	Večerní slunce spočívá na řece .
<input type="checkbox"/>	Fajn život	zde světlo , dostatečnou vlhkost vzduchu a dobře tu zimují	Zimní zahrada navazuje na pracovnu .
<input type="checkbox"/>	Návštěva s vraždou	, ale zřejmě to v něm vyprovokovalo to čtení .	Celá hra je o šikanování .

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

Hledat v korpusu

Korpus:

syn2015

--celý korpus--



Typ dotazu:

CQL

[Vložit tag](#) | [Vložit 'within'](#) | [Klávesnice](#) | [Předchozí dotazy](#)

Dotaz:

```
[afun="Pnom"&pos="P"&case="7"]
```

[Předchozí dotazy lze zobrazit také pomocí šipky dolů](#) ([další tip](#))

Výchozí atribut: word

[\(popis morf. značek\)](#)

► **Specifikovat kontext**

► **Vybrat texty**

Hledat

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

1. Najděte všechna slova se syntaktickou funkcí **Sb**.
 - a) Odfiltrujte slovesa v určitém tvaru (alespoň Vp, VB, Vi, Vs).
 - b) Zjistěte frekvenci slovních druhů.
 - c) U substantiv a zájmen s funkcí **Sb** zjistěte frekvenci pádů; jsou jiné pády než nominativ v pořádku?

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

2. Najděte všechna slova se syntaktickou funkcí **Obj**.

- a) Odfiltrujte slovesa v určitém tvaru (alespoň Vp, VB, Vi, Vs).
- b) Zjistěte frekvenci slovních druhů.
- c) U substantiv a zájmen s funkcí **Obj** zjistěte frekvenci pádů.
- c) Najděte slovesa v určitém tvaru (Vp, VB, Vi, Vs) s funkcí Obj. Jaké věty reprezentují, na čem jsou tyto věty závislé?

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

3. Najděte substantiva s funkcí **Pnom** v nominativu a instrumentálu, v korpusu SYN2015 a jeho subkorpusech beletrie, odborné literatury a publicistiky.

Odfiltrujte slovesa v určitém tvaru (alespoň Vp, VB, Vi, Vs).

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

3. Najděte substantiva s funkcí **Pnom** v nominativu a instrumentálu, v korpusu SYN2015 a jeho subkorpusech beletrie, odborné literatury a publicistiky.

Odfiltrujte slovesa v určitém tvaru (alespoň Vp, VB, Vi, Vs).

[pos="N" & afun="Pnom" & case="1|7"]

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

3. Najděte substantiva s funkcí **Pnom** v nominativu a instrumentálu, v korpusu SYN2015 a jeho subkorpusech beletrie, odborné literatury a publicistiky.

Odfiltrujte slovesa v určitém tvaru (alespoň Vp, VB, Vi, Vs).

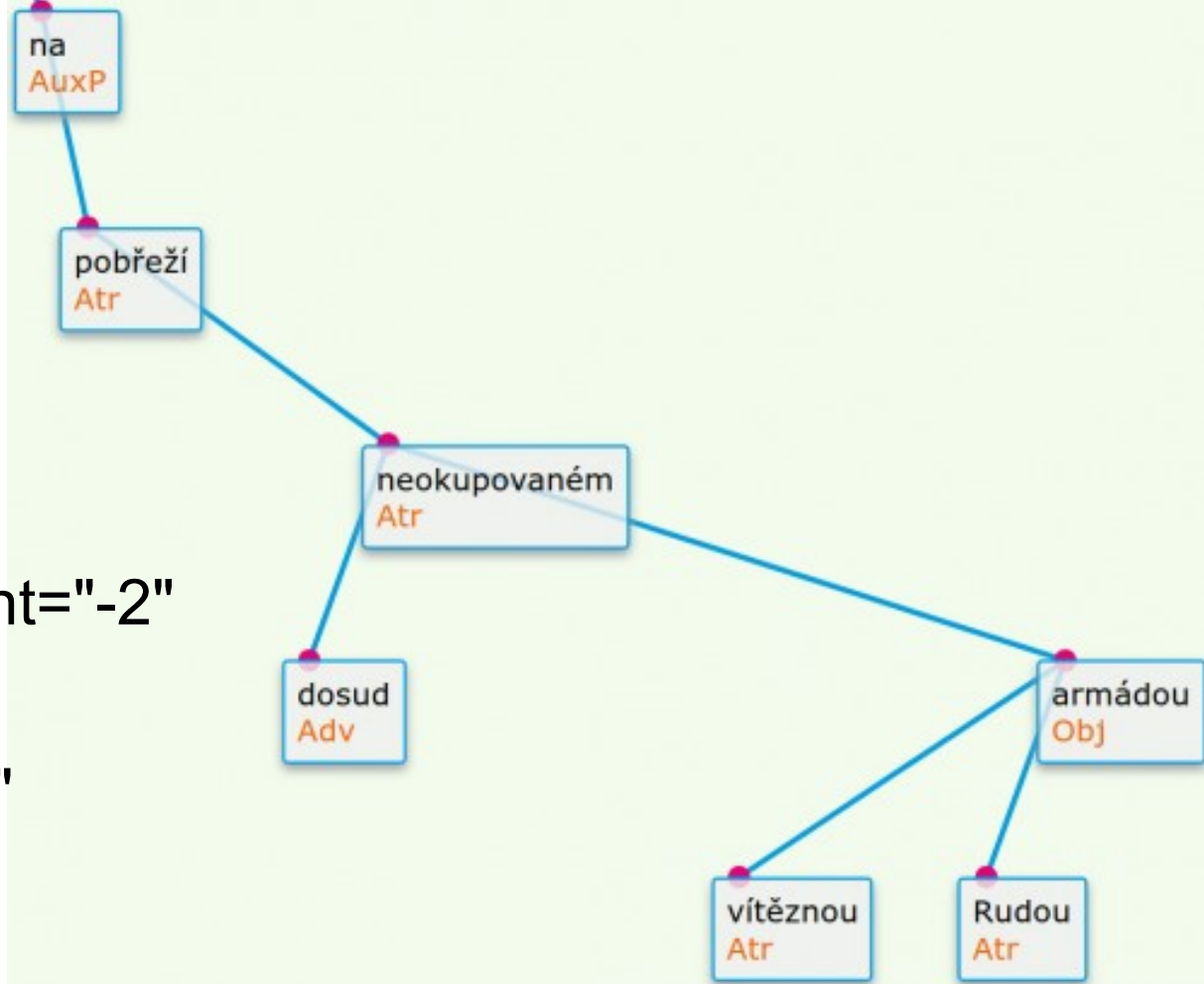
- a) Jaký je poměr frekvencí nominativu a instrumentálu v jednotlivých subkorpusech?
- b) Navrhnete hypotézu, proč tomu tak je?
- c) Jaké jsou nejčastější tvary slov v této funkci, v nominativu a v instrumentálu?
- d) Je poměr u zájmen stejný?

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

- vzdálenost

parent="-2,,

parent="\+2"



- *pobřeží* parent="-1"
- *neokupovaném* parent="-2"
- *vítěznou* parent="\+2"
- *Rudou* parent="\+1"
- *armádou* parent="-3"

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

Chceme-li tak například najít všechny řadové číslovky závislé na těsně následujícím tokenu, zadáme následující příkaz:

```
[parent="\+1" & tag="Cr.+"]
```

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

Pro vyhledání řadových číslovek závislých na jakémkoli následujícím tokenu (tedy v libovolné vzdálenosti), použijeme dotaz o trochu složitější:

```
[parent="\+[0-9]+" & tag="Cr.+"]
```

Hledáme totiž slova, která zároveň splňují tyto podmínky (&):

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

Pro vyhledání řadových číslovek závislých na jakémkoli následujícím tokenu (tedy v libovolné vzdálenosti), použijeme dotaz o trochu složitější:

```
[parent="\+[0-9]+" & tag="Cr.+"]
```

Hledáme totiž slova, která zároveň splňují tyto podmínky (&):

1. morfologická značka začíná na písmena C (číslovka) a r (řadová), následují libovolné znaky: tag="Cr.+"
2. atribut parent začíná znakem plus \+, následuje libovolný počet čísel 0 až 9 (minimálně jedno): parent="\+[0-9]+"

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

1. Najděte příklady, kdy substantivum závisí na těsně předcházejícím slově.

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

1. Najděte příklady, kdy substantivum závisí na těsně předcházejícím slově.
2. Najděte příklady, kdy substantivum závisí na těsně následujícím slově.

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

1. Najděte příklady, kdy substantivum závisí na těsně předcházejícím slově.
2. Najděte příklady, kdy substantivum závisí na těsně následujícím slově.
3. Najděte případy, kdy po sobě následují dvě substantiva, z nichž první je závislé na druhém, a shodují se v pádu.

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

1. Najděte příklady, kdy substantivum závisí na těsně předcházejícím slově.
2. Najděte příklady, kdy substantivum závisí na těsně následujícím slově.
3. Najděte případy, kdy po sobě následují dvě substantiva, z nichž první je závislé na druhém, a shodují se v pádu.
4. Najděte případy, kdy po sobě následují dvě substantiva, z nichž druhé je závislé na prvním, a shodují se v pádu (ne nominativ, ne genitiv)

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

1. Najděte příklady, kdy substantivum závisí na těsně předcházejícím slově.
2. Najděte příklady, kdy substantivum závisí na těsně následujícím slově.
3. Najděte případy, kdy po sobě následují dvě substantiva, z nichž první je závislé na druhém, a shodují se v pádu.
4. Najděte případy, kdy po sobě následují dvě substantiva, z nichž druhé je závislé na prvním, a shodují se v pádu (ne nominativ, ne genitiv)
1:[pos="N" & case="3|4|5|6|7"] 2:[pos="N" & parent="-1"] & 1.case=2.case

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

1. Najděte dvě po sobě jdoucí adjektiva, z nichž první je závislé na předcházejícím slově, druhé na následujícím slově.

VYHLEDÁVÁNÍ PODLE ŘÍDÍCÍHO SLOVA

1. Najděte dvě po sobě jdoucí adjektiva, z nichž první je závislé na předcházejícím slově, druhé na následujícím slově.

[pos="A" & parent="-1"] [pos="A" & parent="\+1"]

NEPROJEKTIVITA

A tak se potřebujeme o něco opřít , abychom ten nápor

Něco vytušili a **ted' nás chtěj vylákat** . . .

Prolítnu sprchou a **pak si můžeme dát skleničku vína** ,

Je o celých osm let mladší než já a osudné okamžiky

Ačkoli se to snažil nedat najevo , předpokládal , že odpověď

Když ho potřebuju vyfotit bez cigarety , tak mám někdy

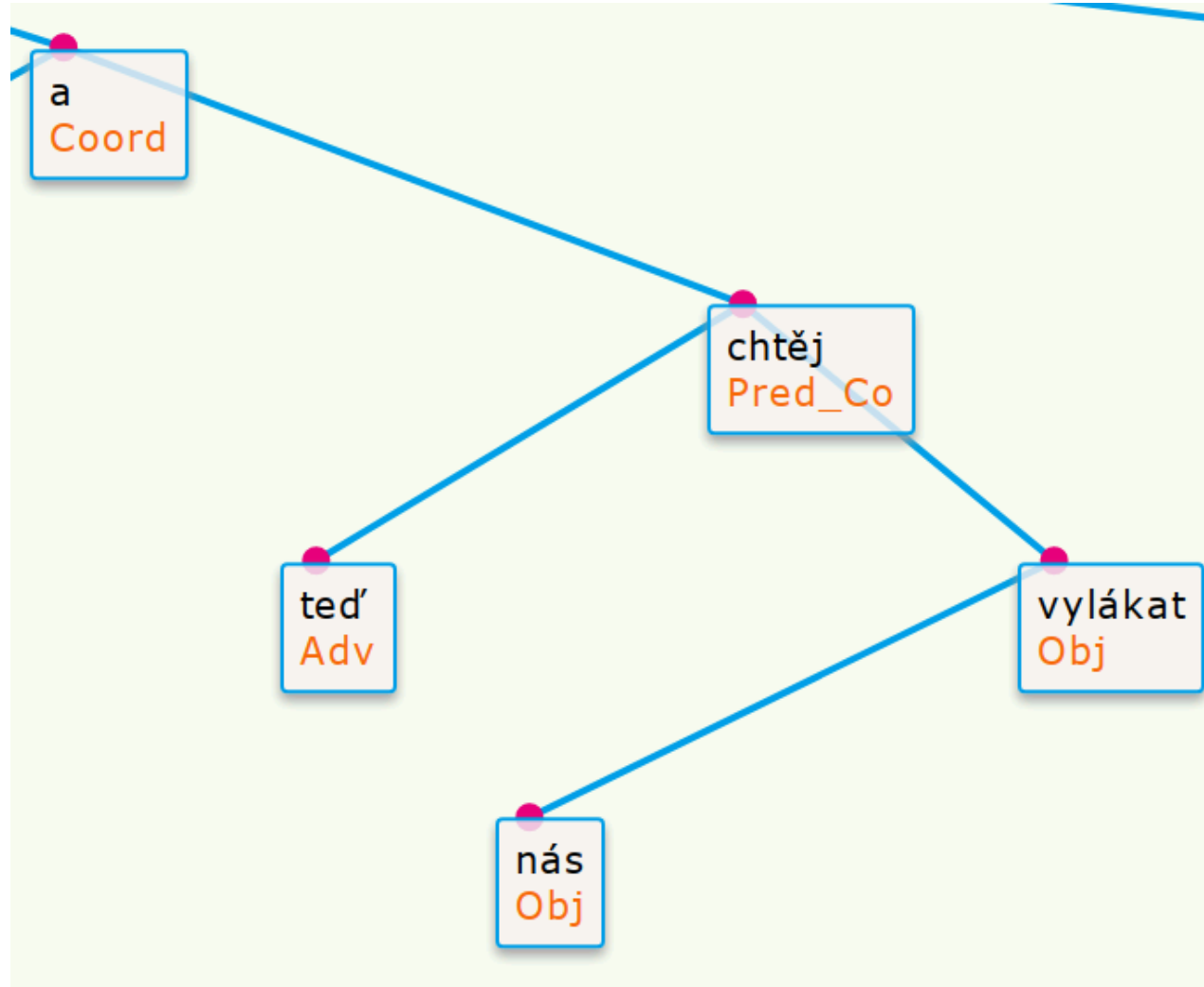
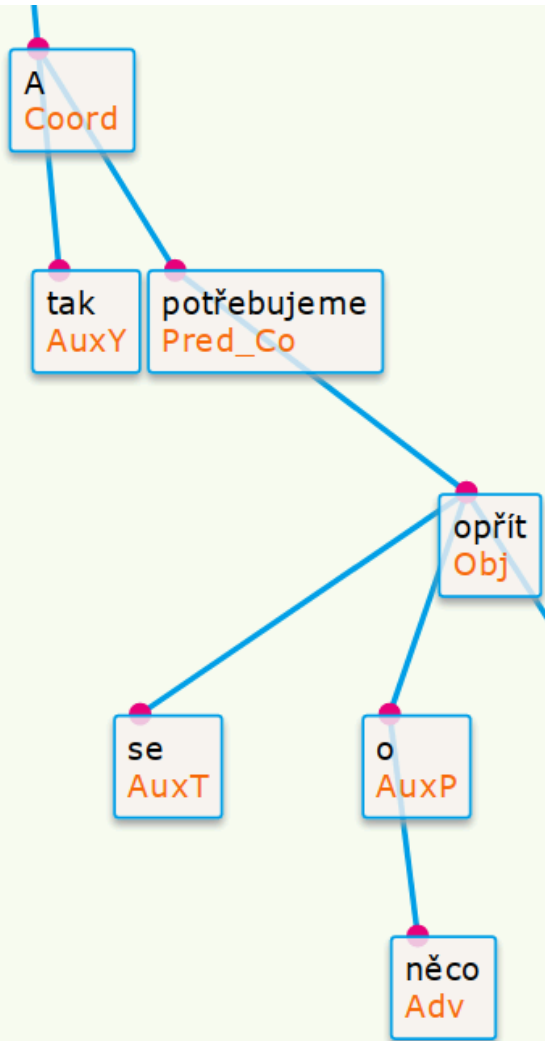
Zápisník končí , **oblíbený doplněk má nahradit služba Keep**

Dobře , **že Max se může postarat o nás** .

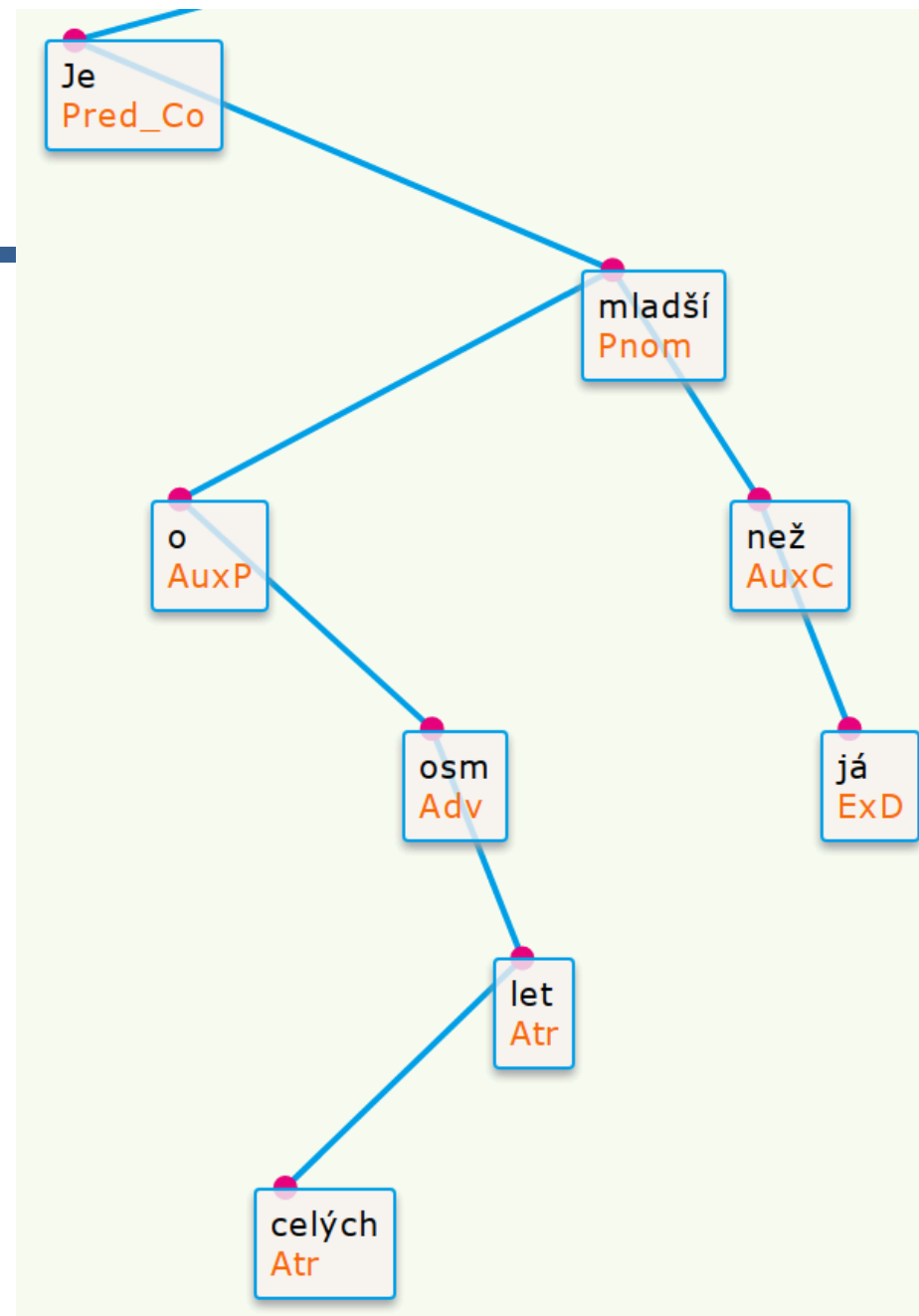
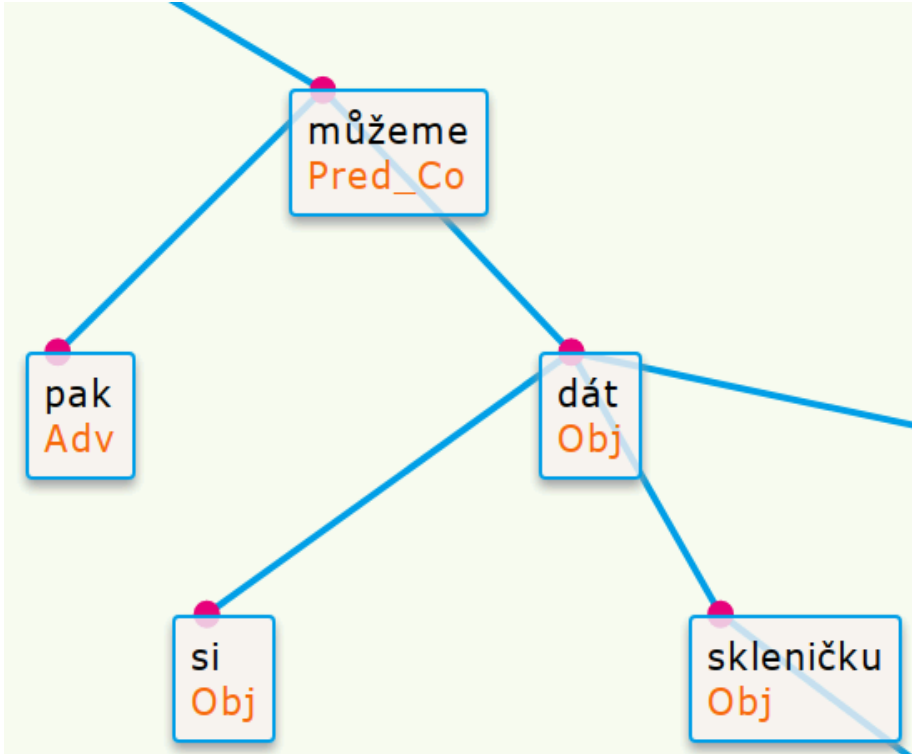
To se lehkou řekne , **ale jak to mám asi zvládnout ?**

Ale o něm mi zakázala mluvit .

NEPROJEKTIVITA



NEPROJEKTIVITA



NEPROJEKTIVITA

Možná řešení:

1. Obecnější, ale vyžaduje „manuální“ kontrolu:

Najít případy, kdy se závislosti dvou po sobě jdoucích tokenů kříží

NEPROJEKTIVITA

Možná řešení:

1. Obecnější, ale vyžaduje „manuální“ kontrolu:

Najít případy, kdy se závislosti dvou po sobě jdoucích tokenů kříží

2. Zacílenější, ale pro nalezení všech případů nutno zadat více dotazů:

Najít případy, kdy slovo (nebo více slov!) uprostřed nezávisí ani na předcházejícím, ani na následujícím slově, přitom předcházející závisí na následujícím či naopak

NEPROJEKTIVITA

Řešení 1 (například):

CQL Dotaz:

```
[pos="[NAPCVD]"&parent="\+[2345]"][pos="[N  
APCVD]"&parent="-[2345]"]
```

*Byl detekován možný nesoulad mezi zvoleným a faktickým typem dotazu.
Chcete pokračovat? OK*

Z výsledků nutno (ručně či vhodně zvolenými filtry) odstranit případy, kdy řídicí slova nejsou zákl. slov. druhy...

NEPROJEKTIVITA

Řešení 2 (například):

CQL Dotaz:

```
[pos="[NAPCDV]"&parent!="\+1"]
```

```
[pos="[NAPCDV]"&parent!="\+1"&parent!="-1"]
```

```
[pos="[NAPCDV]"&parent="-2"]
```

Byl detekován možný nesoulad mezi zvoleným a faktickým typem dotazu. Chcete pokračovat? OK

Takto vyhledáme jen jeden typ (neprojektivní vztah na vzdálenost dvou tokenů, „vpravo“). Budeme jich potřebovat najít více...

NEPROJEKTIVITA

Funkční slova (předložky, spojky) v závislostní struktuře zamlžují vztah mezi plnovýznamovými slovy...

Řešení: atribut **eparent** („efektivní“, skutečný rodič)

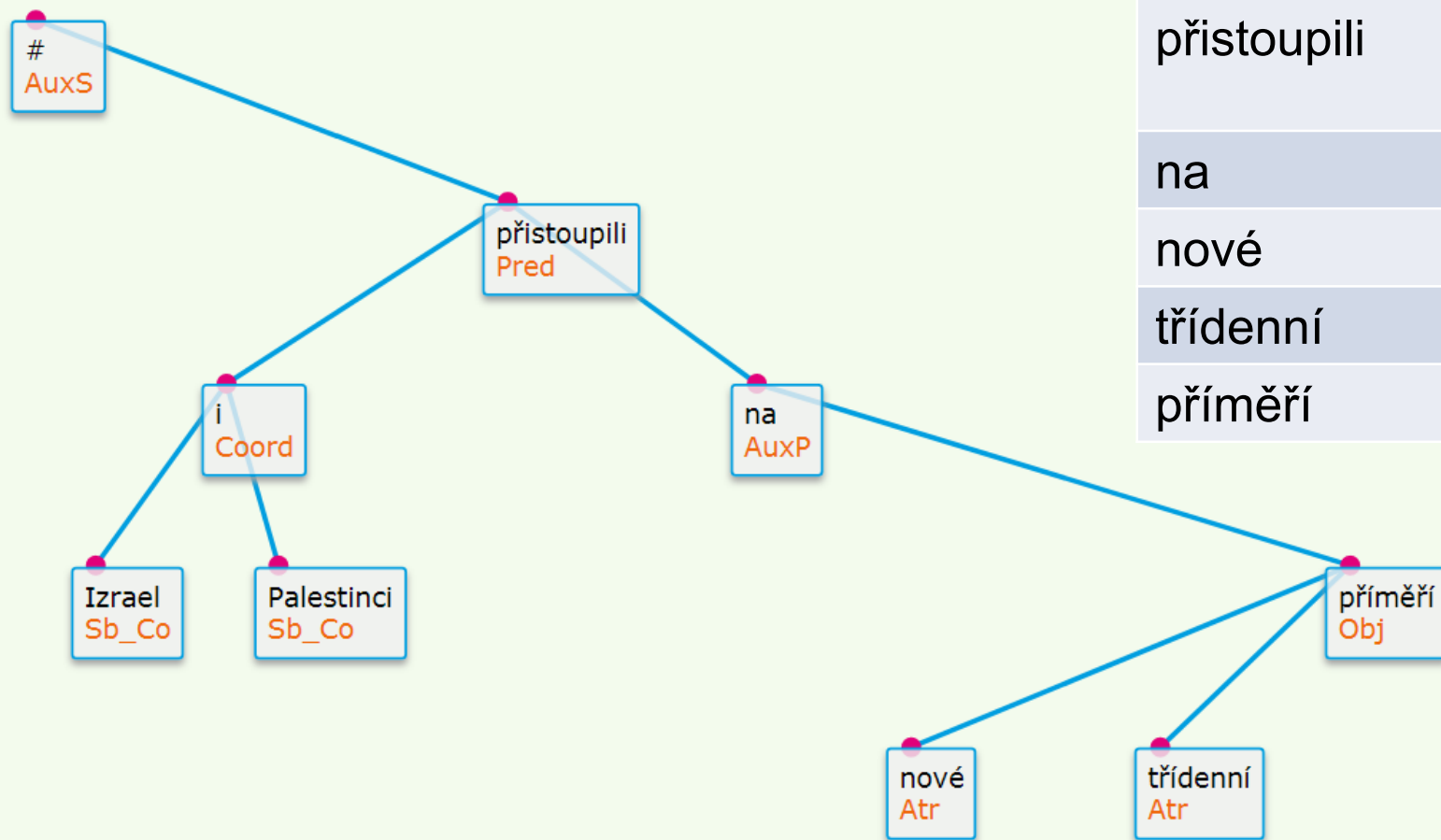
Pouze u plnovýznamových slov (pos="[NAPCVD]")

Odkazuje na nejbližší nadřízené **plnovýznamové** slovo

Hodnoty stejné jako u atributu **parent**

NEPROJEKTIVITA

Izrael i Palestinci přistoupili na nové třídenní příměří



word	eparent
Izrael	+3
i	
Palestinci	+1
přistoupili	0
na	?
nové	?
třídenní	?
příměří	?

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

1. Najděte slova s funkcí **AuxV**.

Jaké slovní tvary jsou v této funkci nejčastější?

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

1. Najděte slova s funkcí **AuxV**.

Jaké slovní tvary jsou v této funkci nejčastější?

2. Najděte slova s funkcí **AuxT** nebo **AuxR**.

Je z kontextu patrný rozdíl mezi těmito funkcemi?

Jaké jiné syntaktické funkce má v korpusu reflexivum „se“?

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

3. Najděte slova s funkcí **Coord**, potom slova s funkcí **Apos**.
Jaká konkrétní slova/tokeny jsou v těchto funkcích nejčastější?
4. Najděte slova, jejichž syntaktická funkce má příponu **_Co**
Která syntaktická funkce je nejčastěji koordinovaná?
5. Najděte slova, jejichž synt. funkce má příponu **_Apos**
Která syntaktická funkce, který slovní druh je nejčastěji v apozičním vztahu?

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

Atribut p_pos [p_pos="N"]

1. Najděte předložky závislé na substantivu. Zjistěte, která substantiva jsou nejčastěji rozvířá kterými předložkami.
2. Najděte předložky závislé na číslovce. Zjistěte, která předložka je takto závislá nejčastěji.

VYHLEDÁVÁNÍ POMOCÍ SYNT. FUNKCE

Atribut `p_case` [`p_case="7"`]

- Najděte substantiva v instrumentálu závislá na vlevo stojícím substantivu, které není v instrumentálu. Zjistěte, která substantiva jsou nejčastěji rozvitá substantivem v instrumentálu.
- Najděte substantiva v genitivu závislá na číslovce, která není ani v genitivu, ani v nominativu či akuzativu. Pokud je výsledek jakžtakž správný, která číslovka je v takovém spojení nejčastější?



Koordinace a apozice

- Najděte slova s funkcí **Coord**, potom slova s funkcí **Apos**.
 - Jaká konkrétní slova/tokeny jsou v těchto funkcích nejčastější?
- Najděte slova, jejichž syntaktická funkce má příponu **_Co**
 - Která syntaktická funkce je nejčastěji koordinovaná?
- Najděte slova, jejichž synt. funkce má příponu **_Apos**
 - Která syntaktická funkce, který slovní druh je nejčastěji v apozičním vztahu?





Atribut **e**parent

- Najděte všechna zájmena, která jsou (přímo či nepřímo) závislá na slově nacházejícím se o pět slov dále (vlevo či vpravo).
- Najděte případy, kdy substantivum přímo závisí na slově vlevo od něj, nepřímo na plnovýznamovém slově vpravo. Zobraďte si pár takových struktur.





Atributy **ep_lemma**, **ep_afun**

- Najděte případy, kdy substantivum závisí (přímo či nepřímo) na substantivu *touha*
- Najděte případy, kdy substantivum závisí (přímo či nepřímo) na slovese v infinitivu, které má funkci **Sb**.





Morfologická desambiguace

Slovo **se**

Navrhněte několik jednoduchých pravidel, která na základě užšího či širšího kontextu rozliší, zda je tvar *se* reflexivum nebo vokalizovaná předložka.

Slovo **je**

Navrhněte několik jednoduchých pravidel, která na základě užšího či širšího kontextu rozliší, zda je tvar *je* sloveso nebo osobní zájmeno.



Literatura

Nový encyklopedický slovník češtiny online:

<https://www.czechency.org/>

hesla: Formální gramatika

Wiki ČNK

<https://wiki.korpus.cz/doku.php/start>