

Nainstalovat Tred!

<http://ufal.mff.cuni.cz/tred/>



Korpusy a lingvistické nástroje

Daniel Zeman

zeman@ufal.mff.cuni.cz

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



informatika

+

=

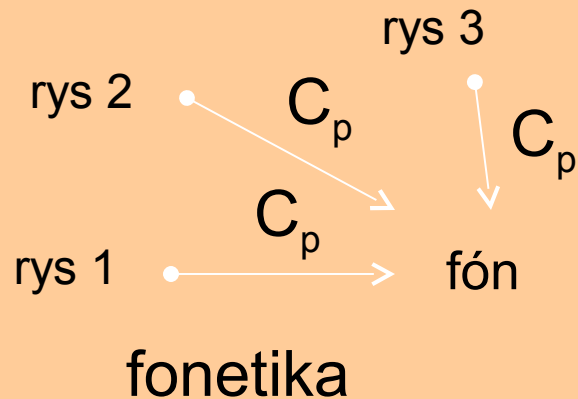
lingvistika



Teoretické ukotvení

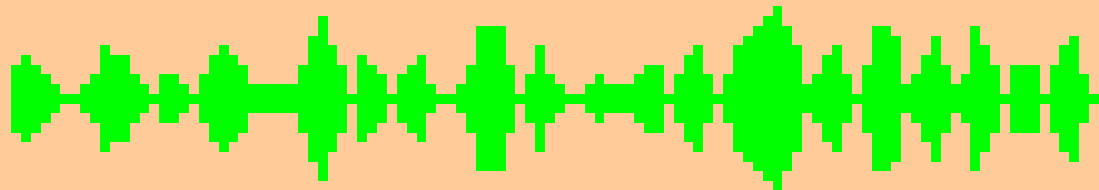
- Funkční generativní popis jazyka (FGP)
 - *Petr Sgall et al., 80. léta 20. století*
- Systém rovin
- Závislostní syntax
- Valence

Fonetická rovina

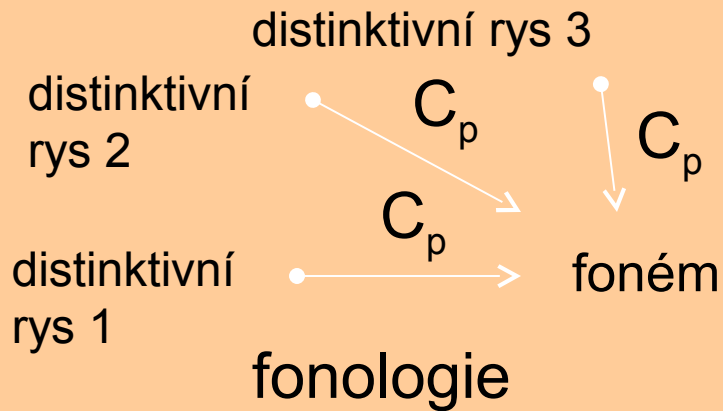


z	b	a	ŋ	k	o	u
---	---	---	---	---	---	---

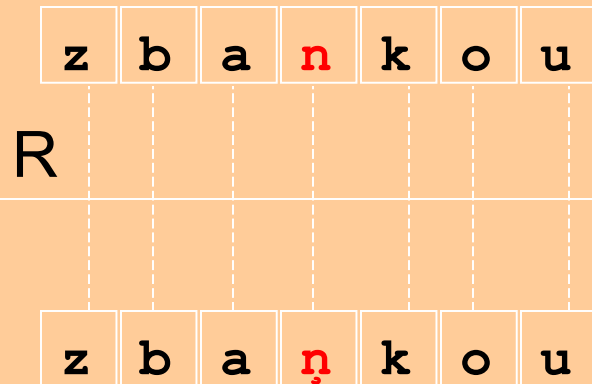
akustický
signál



Fonologická rovina



fonetika



Pravopisná rovina

fonologie

fonémy

j	a	b	l	o	ň	e
---	---	---	---	---	---	---

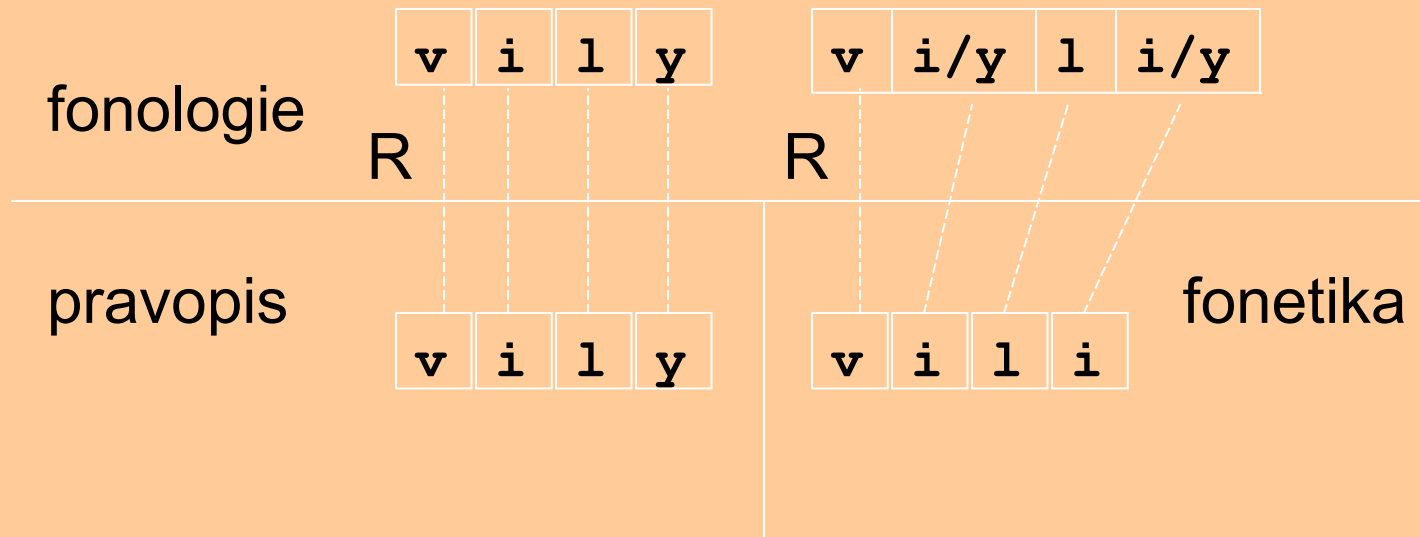
R

pravopis

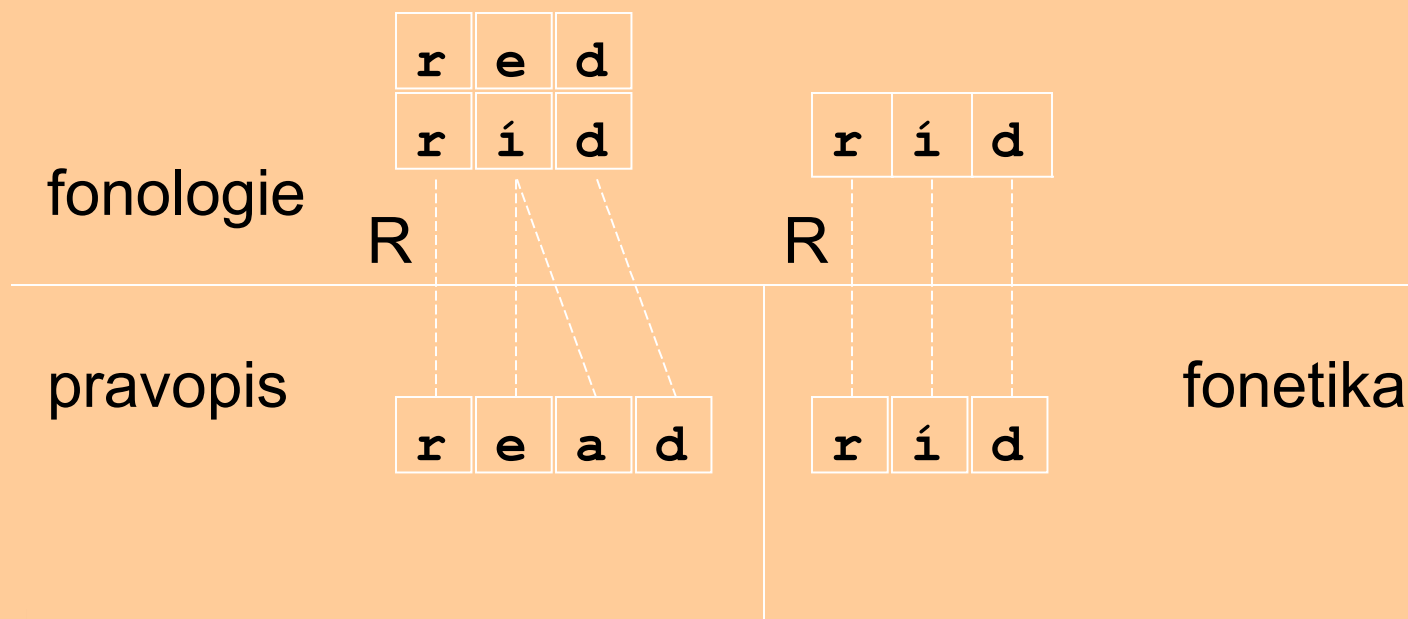
j	a	b	l	o	n	ě
---	---	---	---	---	---	---

grafémy

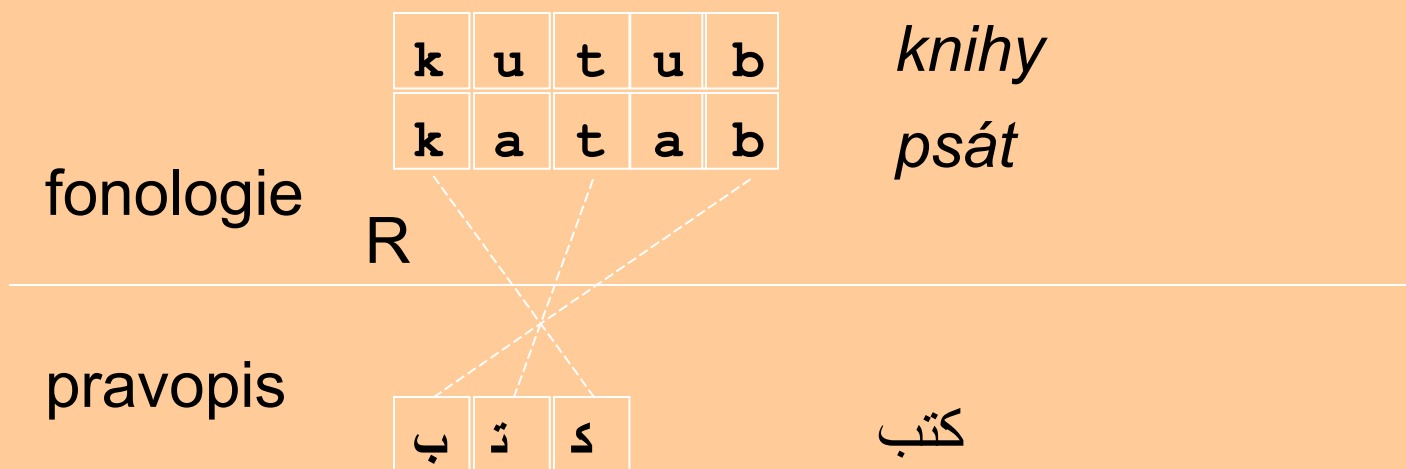
Stejné znění, různý zápis



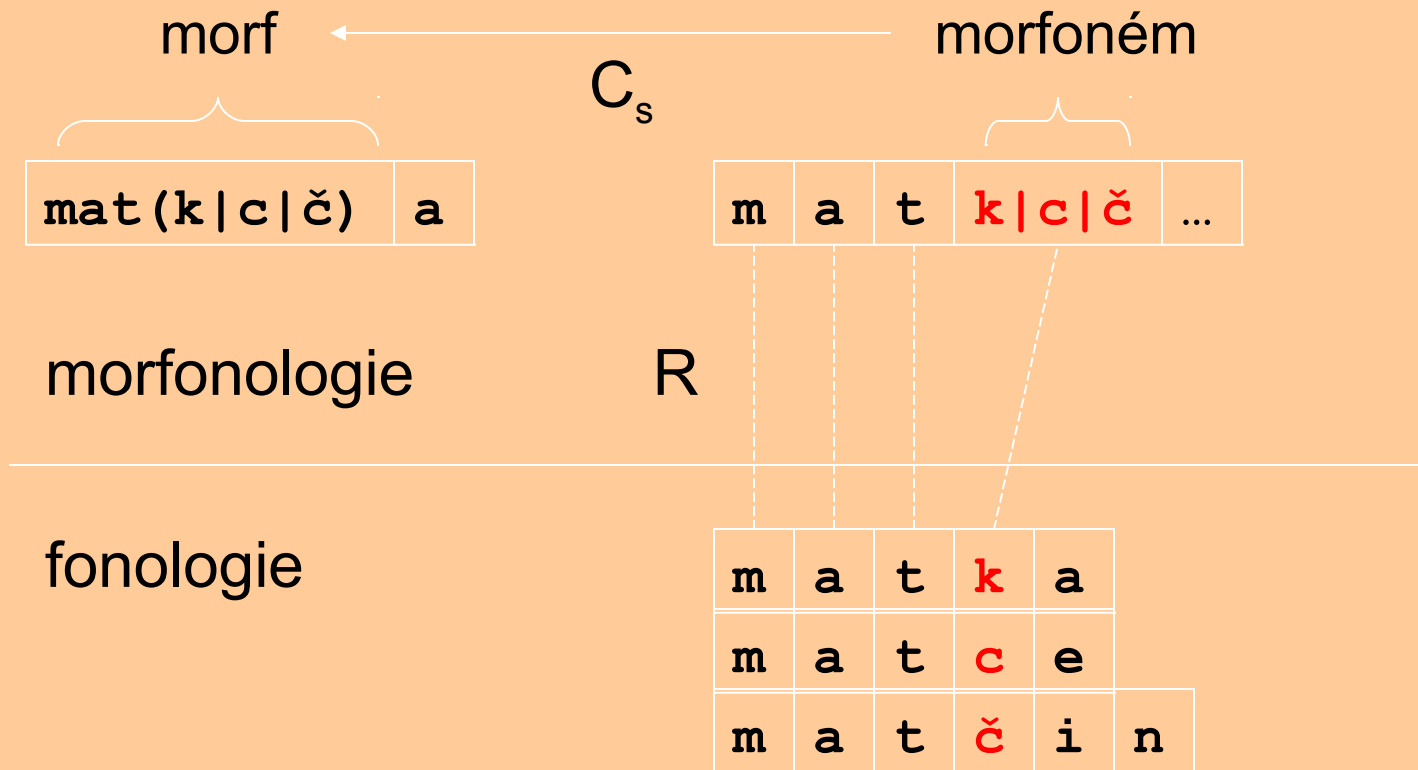
Opačný případ: stejný zápis, jiná výslovnost



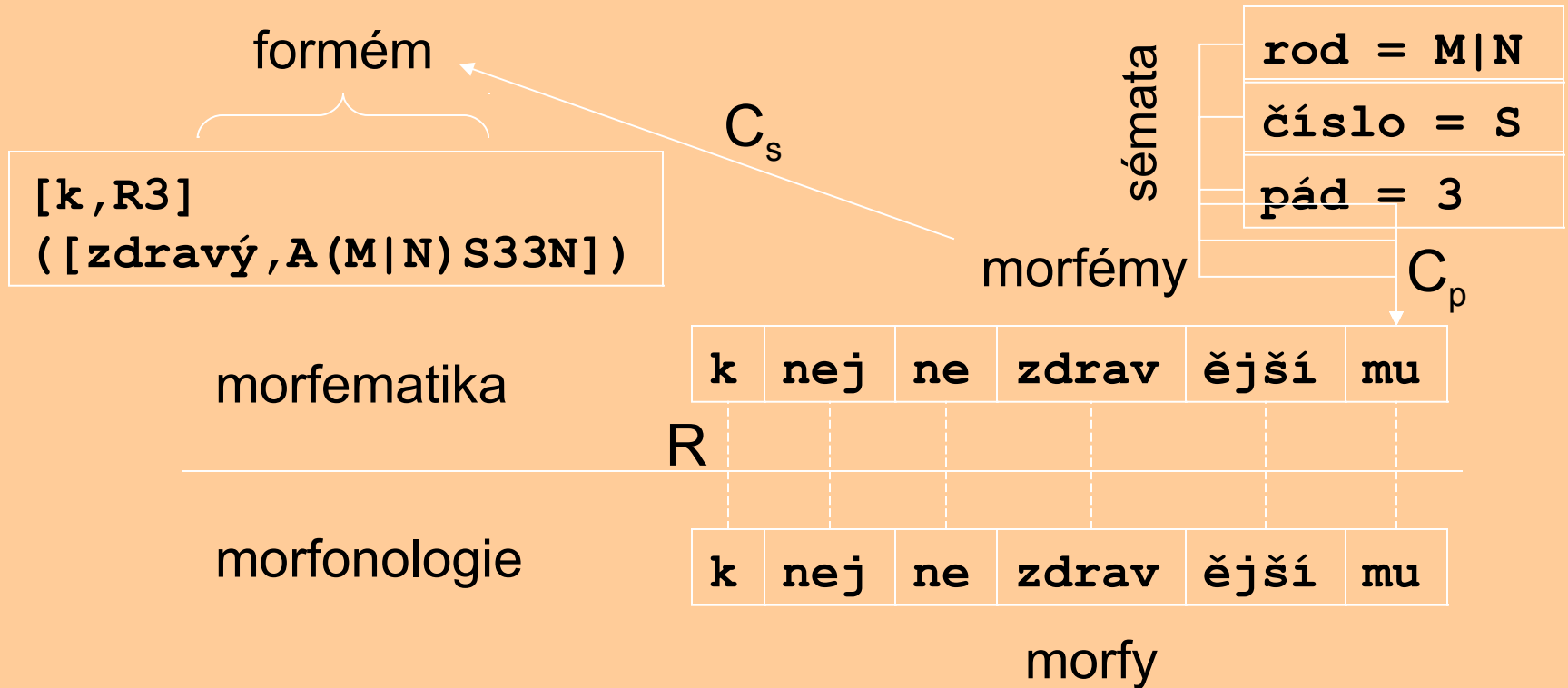
Opačný případ: stejný zápis, jiná výslovnost



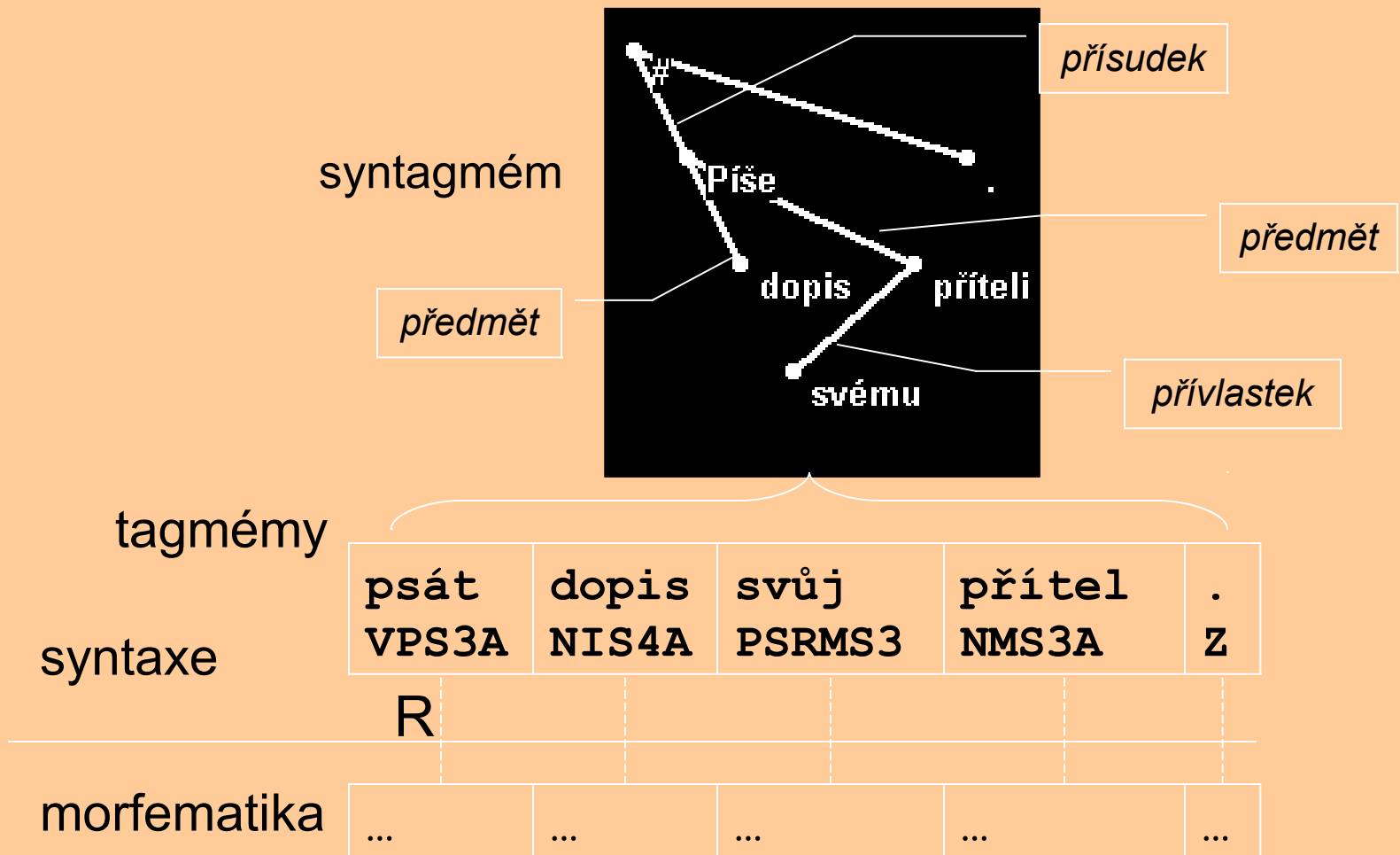
Morfonologická rovina



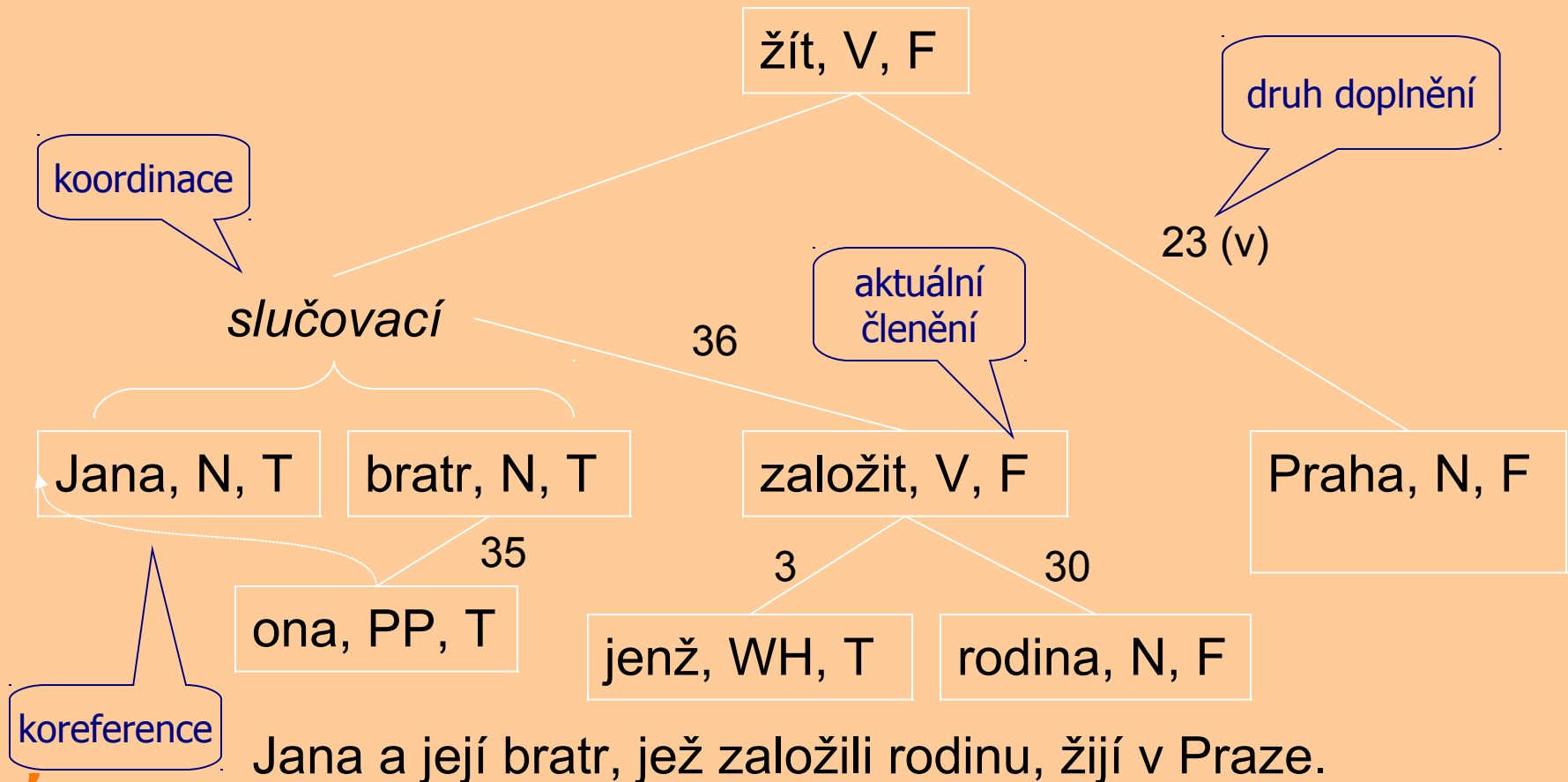
Morfematická rovina



Povrchově syntaktická rovina



Tektogramatická rovina



Rovina diskurzu

- Vstup:
 - posloupnost vět (propozic) patřících k sobě (např. jeden článek) = diskurz
- Výstup:
 - diskurz s vyřešenými mezivětnými vztahy
- Obsah:
 - např. anafora, katafora:
„Pavel přišel po desáté. Dala jsem mu večeři.“



Pragmatická (logická) rovina

- Vstup:
 - hloubková struktura věty (propozice)
- Výstup:
 - logická forma, která může být vyhodnocena (pravda/nepravda)
- Obsah:
 - přiřazení objektů reálného světa uzlům větné struktury
 - mimolingvistický obsah: znalost světa („oči barvy nebe“)
 - kvantif. („mnoho knih čte málo lidí“ × „mnoho lidí čte málo knih“)
 - například (já/Sg/Pat/t (vidět(Past/Pred/t) babička/Sg/Ag/f) ~ vidět(babička-BN[SSN:...], Němcová[SSN:...])_[čas:před 13.12.2013]
_[místo:mezi (50°20'00"N15°30'00"E, 51°00'00"N16°30'20"E)]

Praxe: 3 hlavní roviny

- **tektogramatická**
 - hloubkový slovosled (aktuální členění)
 - koordinace
 - koreference, doplnění vypuštěných členů, spojování členů
- **analytická** (syntaktická)
 - posloupnost dvojic → stromová struktura
- **morfologická**
 - slovní tvar → dvojice [lemma, značka]
- **předzpracování**
 - posloupnost znaků (písmen) → posloupnost slov (a interpunkce)

Doprovodný web

<http://ufal.mff.cuni.cz/daniel-zeman/nastroje-ufal>

(http://bit.ly/nastroje_ufal)



Webové nástroje

- Hlavní výpočet na vzdáleném serveru
- Ovládáte prostřednictvím webového prohlížeče
- U sebe nic dalšího neinstalujete
- Minimální nároky na přenositelnost (Windows / Linux...)
- Potřebujete připojení k internetu
- Nevhodné pro velké množství dat

- Většina nástrojů a dat k dispozici i off-line
 - Někdy přenositelné (Java, Perl, ...)
 - Mnohdy nároky na procesor, paměť, disk

Morfologická analýza češtiny

- Vstup:
 - slovo v libovolném tvaru
- Výstup:
 - základní tvar slova (lemma)
 - morfologická značka
 - zakódovaný slovní druh
 - morfologické kategorie
- Homonyma: pro každý význam samostatný výstup!
 - Na kontext (zatím) nehledíme

Morfologické značky

- Smluvená „šifra“
- Jiný korpus ⇒ jiná sada značek?
 - Ajka (Brno) [846]
 - Poziční, Hajičovy (Praha, PDT, ČNK, ČAK...) [4294]
 - CoNLL (vycházejí z pražských, vypadají jinak)
 - CoNLL 2006 a 2007 [5150]
 - CoNLL 2009 [5150]
 - Multext-East (Orwell, ÚTKL) [1458]
 - Pražský mluvený korpus
 - dlouhé [12359]
 - krátké [244]

Brněnské značky (Ajka)

1. *Pekař*

- k1gMnSc1

2. *peče*

- k5eAaImIp3nS

3. *housky*

- k1gFnPc4

4. *v*

- k7c6

5. *pekárně*

- k1gFnSc6

6. .

Pražské poziční značky (Hajič)

1. *Pekař*
2. *peče*
3. *housky*
4. *v*
5. *pekárně*
6. *.*

- NNMS1 - - - - - A - - - -
- VB - S - - - 3P - AA - - -
- NNFP4 - - - - - A - - - -
- RR - - 6 - - - - - - - - -
- NNFS6 - - - - - A - - - -
- Z : - - - - - - - - - - -

CoNLL 2006 (podle pražských)

1. *Pekař* • N N Gen=M | Num=S | Cas=1 | Neg=A
2. *peče* • V B Num=S | Per=3 | Ten=P | Neg=A | Voi=A
3. *housky* • N N Gen=F | Num=P | Cas=4 | Neg=A
4. *v* • R R Cas=6
5. *pekárně* • N N Gen=F | Num=S | Cas=6 | Neg=A
6. *.* • Z : _

CoNLL 2009 (podle pražských)

- | | |
|-------------------|--|
| 1. <i>Pekař</i> | • N SubPOS=N Gen=M Num=S Cas=1 Neg=A |
| 2. <i>peče</i> | • V SubPOS=B Num=S Per=3 Ten=P Neg=A Voi=A |
| 3. <i>housky</i> | • N SubPOS=N Gen=F Num=P Cas=4 Neg=A |
| 4. <i>v</i> | • R SubPOS=R Cas=6 |
| 5. <i>pekárně</i> | • N SubPOS=N Gen=F Num=S Cas=6 Neg=A |
| 6. <i>.</i> | • Z SubPOS=: |

Značky Multext-East

1. *Pekař*
2. *peče*
3. *housky*
4. *v*
5. *pekárně*
6. *.*

- Ncmsa - - y
- Vmip3s - an
- Ncfpa
- Spsl
- Ncfsl
- X

Pražský mluvený korpus (dlouhé značky)

- | | |
|-------------------|----------------------|
| 1. <i>Pekař</i> | • 51 různých značek |
| 2. <i>peče</i> | • 105 různých značek |
| 3. <i>housky</i> | • 72 různých značek |
| 4. <i>v</i> | • 7 různých značek |
| 5. <i>pekárně</i> | • 78 různých značek |
| 6. . | |

Převody mezi sadami značek

- Interset

<https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interiset>

<http://quest.ms.mff.cuni.cz/cgi-bin/interiset/index.pl>

- MorphCon (nadstavba Intersetu)

Petr Pořízka, Markus Schäfer, Daniel Zeman

Převody mezi sadami značek

- Některé sady umí něco, co jiné ne
 - Interpunkce
 - Cizí slova
 - Zkratky
 - Stylové varianty (hovorové, zastaralé atd.)
 - Slovesný vid
 - Detailní rozlišení druhů zájmen, číslovek, příslovcí...
 - Kategorizace vlastních jmen (osoba / místo / organizace)

Interaset

- Univerzální repertoár
 - *rysů (features)*
 - jejich *hodnot (values)*
- Pro každou sadu značek stačí popsat
 - převod sada $\rightarrow$ Interaset („dekódování“)
 - převod Interaset $\rightarrow$ sada („zakódování“)
- Z každé sady do každé pak přes Interaset

Morfologické analyzátory

- Hajičův analyzátor a tagger
<http://ufal-point.mff.cuni.cz/services/morph/>
- Morfo (Hajič / Kolovratník) inflector
<http://quest.ms.mff.cuni.cz/cgi-bin/zeman/morfo/index.pl>
- Ajka (Radek Sedláček, Brno)
<http://nlp.fi.muni.cz/projekty/wwwajka/WwwAjkaSkripty/morph.cgi?jazyk=0>

Homonymie: víceznačné výsledky morfológické analýzy

- **je** = on + PPNS4-3, PPXP4-3 | být + VB-S-3P-AA
- **stát** = stát-1 (země) | stát-2 (na ulici) | stát-3 (milión) | stát-4 (se) | stát-5 (sníh)
- **stane** = stát-4 + VB-S-3P-AA | stanout + VB-S-3P-AA | stan + NNIS5-A
- **hnát** = hnát-1 (noha) | hnát-2 (honit)
- **žena** = žena + NNFS1-A | hnát-2 + VeYS-A
- **kupuje** = kupovat + VB-S-3P-AA | kupovat + VeYS-A
- **růže** = růže + NNFS1-A | NNFS2-A | NNFS5-A | NNFP1-A | NNFP4-A | NNFP5-A

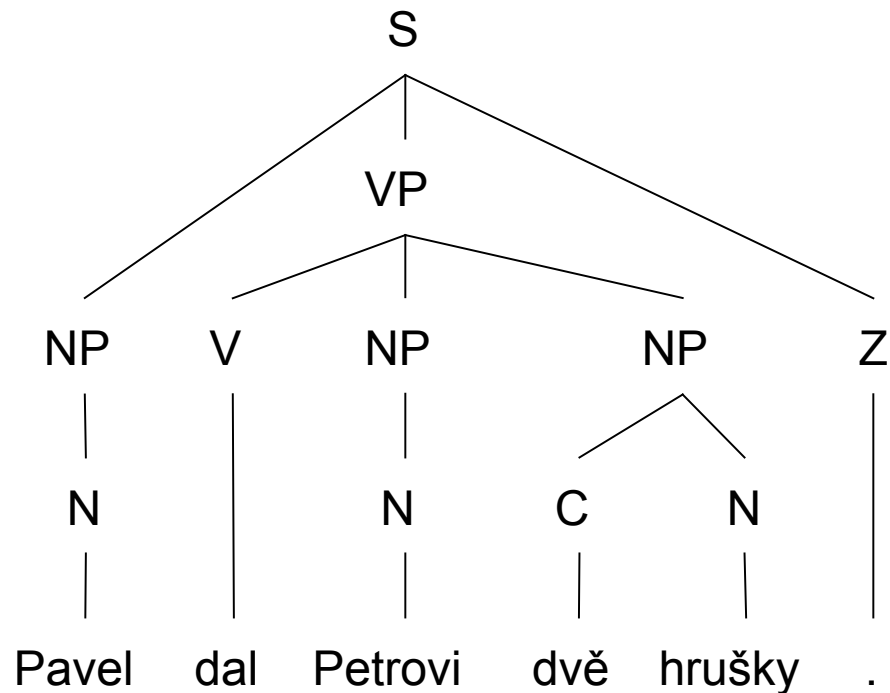
Treebanky

- Zachycují větnou stavbu pomocí *stromů*
- *Uzly* (vrcholy)
- *Hrany* spojují uzly
- Shora dolů: uzel-*rodič* a uzel-*dítě*
- Nejvyšší uzel v hierarchii je *kořen*
- Ostatní uzly mají právě jednoho rodiče
- Uzly bez dětí jsou *listy*



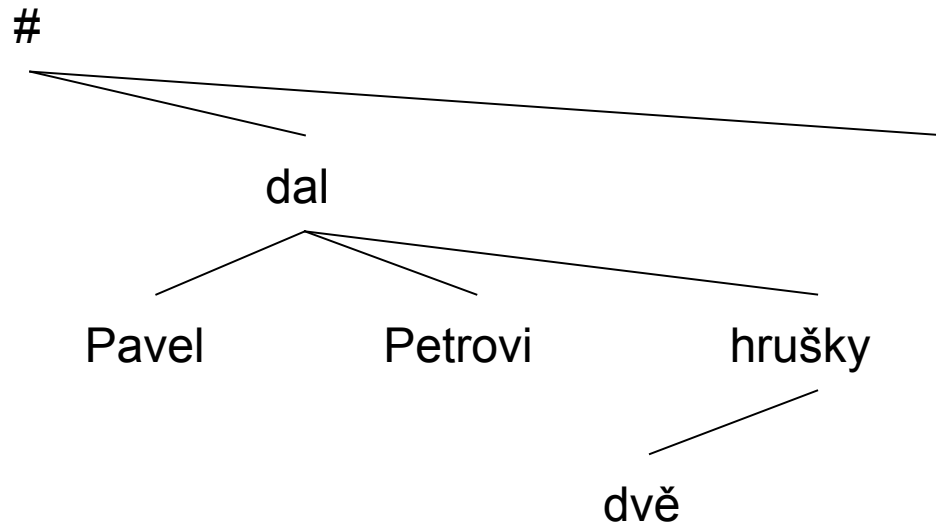
Frázový (složkový) strom

- ((Pavel (dal Petrovi (dvě hrušky)))) .)



Závislostní strom

- $[\#,0] ([dal,2] ([Pavel,1], [Petrovi,3], [hrušky,5] ([dvě,4])), [.,6])$



Pojmenování závislostí (analytické funkce)

/ **AuxS**

dal / **Pred**

. / **AuxK**

Pavel / **Sb**

Petrovi / **Obj**

hrušky / **Obj**

dvě / **Atr**

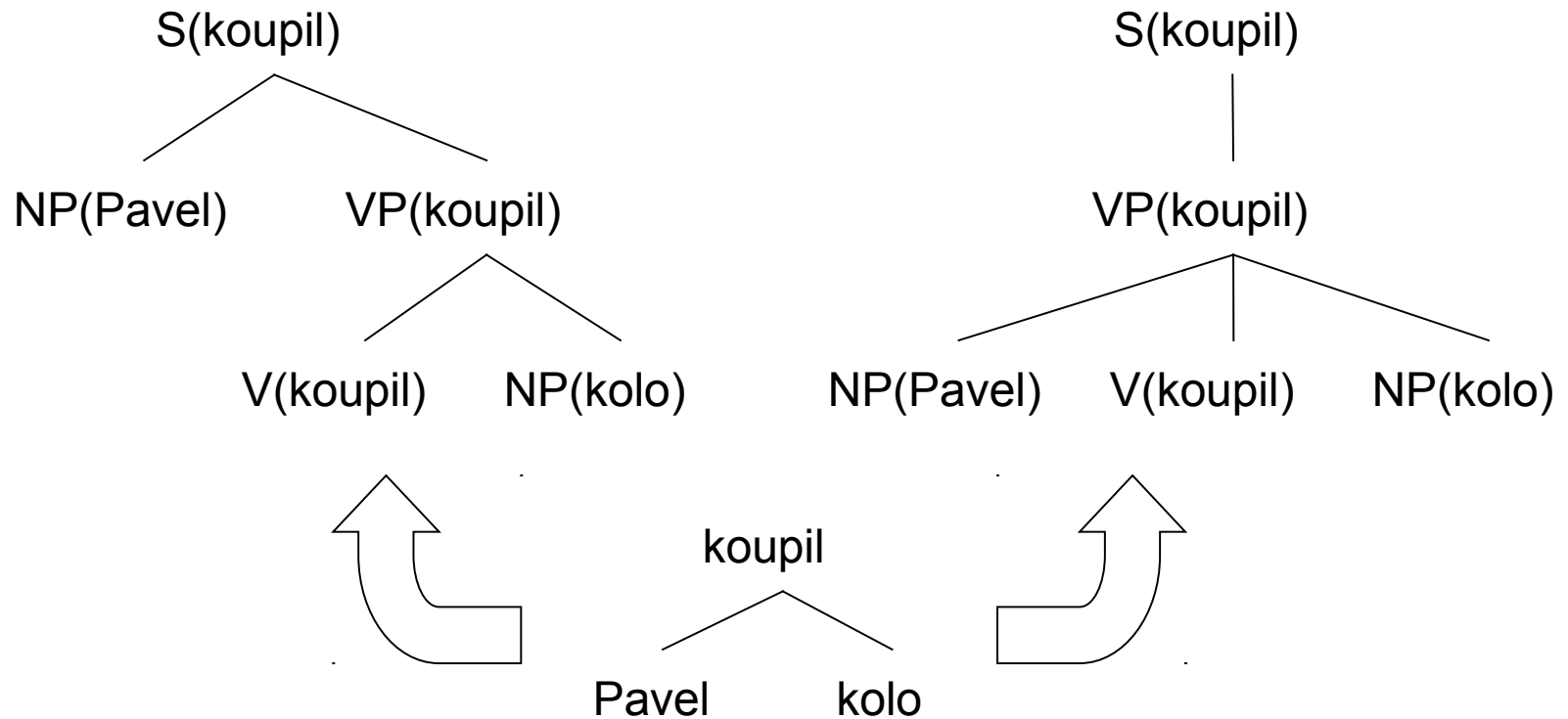
Frázové vs. závislostní stromy

- Frázové (složkové) stromy
 - Ukazují členění věty na fráze a pojmenovávají je.
 - Nezdůrazňují, co je **hlava** fráze, které slovo na kterém závisí.
 - Nemusí obsahovat **funkci**, tj. druh závislosti.
- Závislostní stromy
 - Ukazují závislosti mezi slovy a pojmenovávají je.
 - Nezachycují podobnost tvoření různých částí věty, rekurzi.
 - Nezachycují průběh budování věty, **blízkost** závislých členů na hlavě.
 - Neobsahují **druhy frází**, ty lze leda odhadnout ze značek hlav.

Rozdíly závislostního a frázového modelu

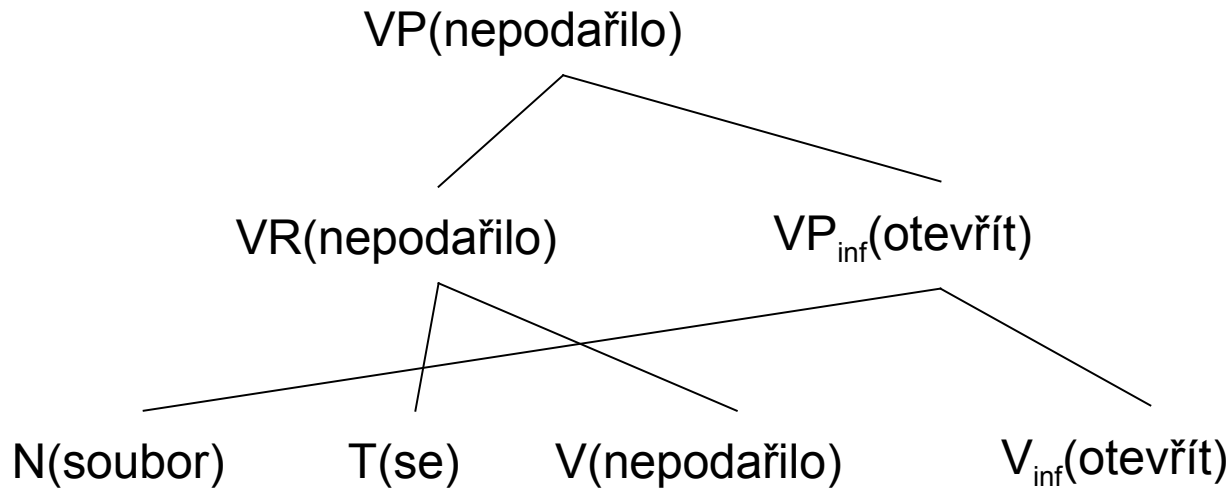
- Chceme převést frázový strom F na závislostní strom Z nebo obráceně.
- Frázový strom neříká, co je hlava fráze.
 - Pro převod $F \rightarrow Z$ potřebujeme výběrovou funkci, která pro každý druh fráze určí hlavu (řídící člen, ostatní členové fráze na něm závisí).
- Závislostní strom neukazuje, jak věta vznikla, ani nutně nepostihuje celé dělení na fráze.
 - Co bylo do věty přidáno dříve a co později?
 - Více frázových struktur vede na stejnou závislostní, převod zpět není jednoznačný.

Příklad



Nespojité fráze

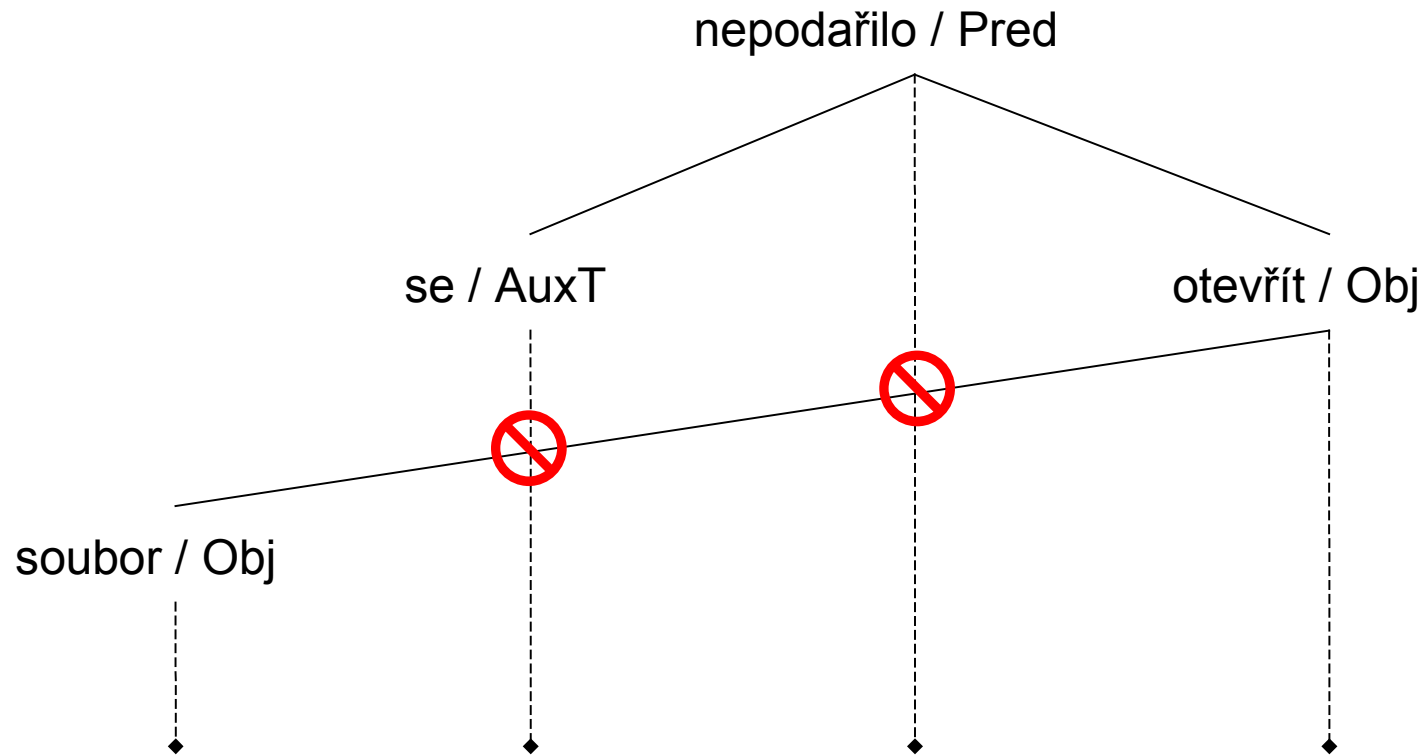
- Frázový strom je pevně spjat se slovosledem (vlastně jen přidává závorky do textu).
- Nespojité fráze v něm nelze zachytit:
(*Soubor (se nepodařilo) otevřít*).



Neprojektivita

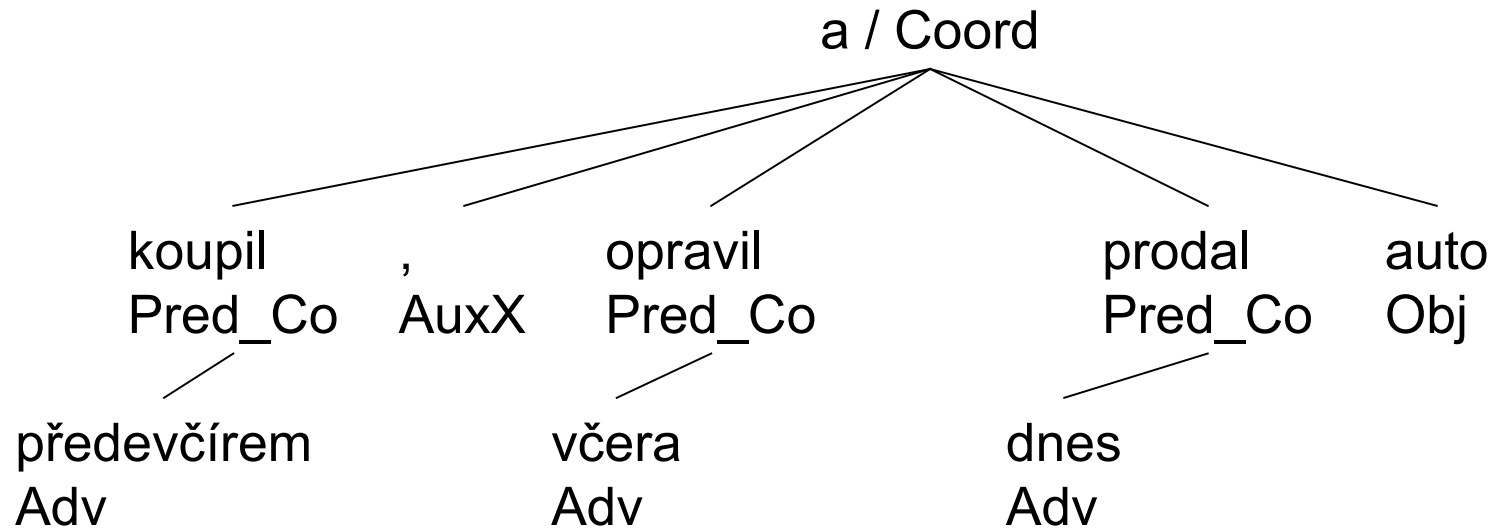
- Závislostní strom nesouvislé fráze umí zachytit.
- Dostaneme podstrom, který nepokrývá souvislý kus věty, je v něm „díra“.
- Tomuto jevu se říká *neprojektivita*.
- Asi 2% slov v PDT jsou zavěšena neprojektivně.
- Asi 20% vět obsahuje alespoň 1 neprojektivitu.

Neprojektivita

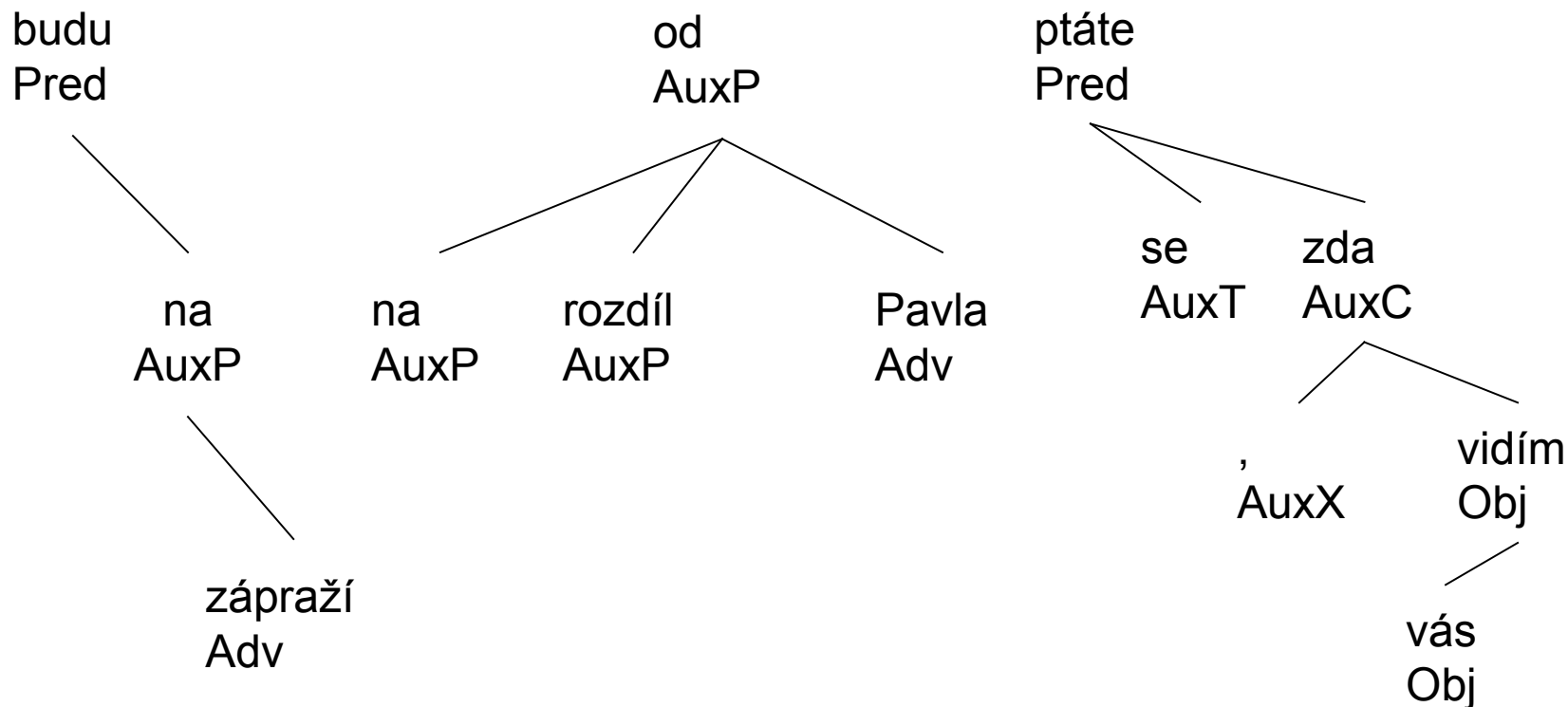


Problém: ne vše je závislost

- Koordinace a apozice
 - Společné rozvití celé koordinace × členy koordinace
 - Pomocné uzly (interpunkce apod.)



Předložkové fráze, vnořené klauze spojkové



Vnořené klauze vztahné

muž
???

představil
Atr

,
AuxX

kterého
Obj

jsem
AuxV

vám
Obj

Vallex

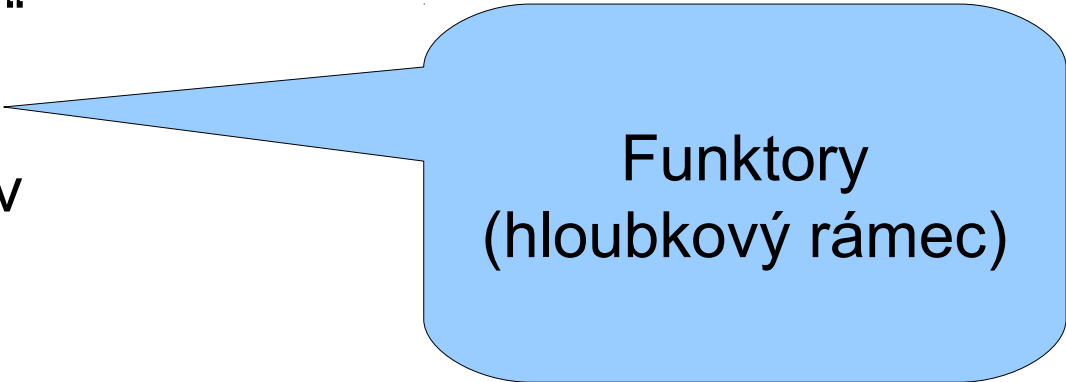
- Elektronický valenční slovník
= slovník vazeb českých sloves
- Z mnoha zdrojů, zejména
 - SSJČ, Slovesa pro praxi
 - PDT a další korpusy

Valenční rámec slovesa

- *budit* ACT PAT (MEANS)
„Budil své děti každé ráno (zpěvem).“
- *budit* ACT PAT (LOC)
= vyvolávat
„Budil (ve mně) podezření.“
- *budit se* ACT
= procítat ze spánku
„Budí se kolem sedmé.“

Valenční rámec slovesa

- *cítit se* ACT PAT
= považovat se za nějakého
„Cítí se jako mladík.“
- *cítit se* ACT MANN
= pociťovat svůj stav
„Cítí se mizerně.“
- *cítit se* ACT PAT
= stačit na něco (idiom)
„Cítí se na tento úkol.“



Funktory
(hloubkový rámec)

Povrchové realizace rámce

- *cítit se* ACT PAT

cítit se Nom Ins

cítit se Nom být+Ins

cítit se Nom jako+Nom

cítit se na+Acc

cítit se být+Adj/Nom

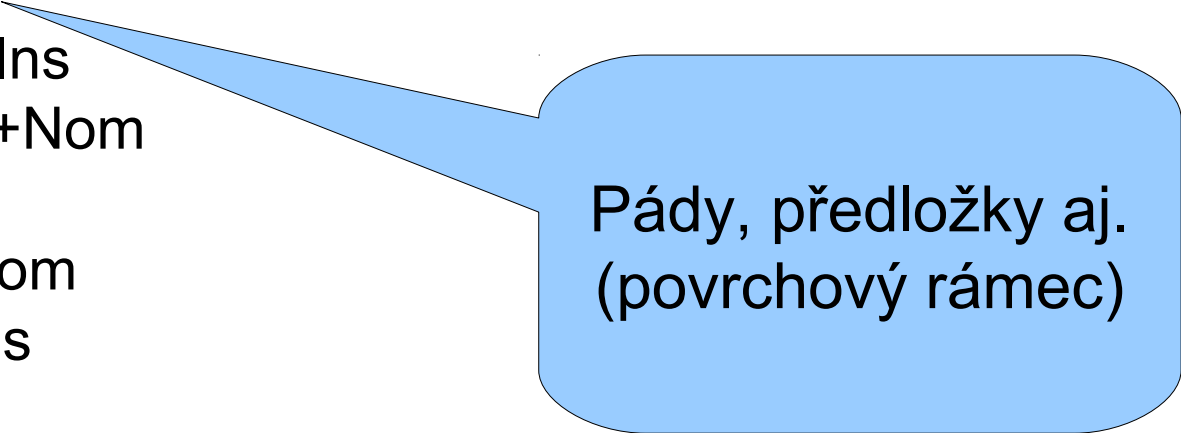
cítit se být+Adj/Ins

cítit se Adj/Nom

cítit se Adj/Ins

- *cítit se* ACT MANN

cítit se Nom ...



Pády, předložky aj.
(povrchový rámec)

Vallex

- Než se do něj podíváme, uděláme si malý test.
- Poznamenejte si na papír nebo do editoru...

nechávat / nechat

- Kolik různých hloubkových rámců (významů) vymyslíte pro sloveso *nechat*?



Reflexiva tantum

- Pouze zvratná slovesa, bez se se nevyskytují.
Příklad: *bát se*
- Kolik takových českých sloves dáte dohromady z hlavy?
- Ostatní funkce zvratného se, si

Povinné doplnění místa

- Která slovesa mají v některém svém významu
 - povinný funktor LOC (odpověď na otázku „kde?“)
- Kolik takových českých sloves dáte dohromady z hlavy?



Valenční genitiv

- Která slovesa mají v některém svém významu
 - povinný předmět v genitivu
 - Příklad: *bát se někoho*
- Kolik takových českých sloves dáte dohromady z hlavy?

Vazba v + akuzativ

- Která slovesa mají v některém svém významu
 - povinný předmět s předložkou v a akuzativem
 - Příklad: ***věřit v něco***
- Kolik takových českých sloves dáte dohromady z hlavy?

Podřadící spojky

- Které podřadící spojky mohou být v češtině valenční?
 - Příklad: *řekl, že*
- Kolik takových českých spojek dáte dohromady z hlavy?



Vidové dvojice (skupiny)

- Slovesa, která jsou současně dokonavá i nedokonavá
- Pouze nedokonavá (bez dokonavého protějšku)
- Pouze dokonavá
- Více než dva tvary (některé dokonavé, některé nedokonavé)

A nyní exkurze do Vallexu...

<http://ufal.mff.cuni.cz/vallex/2.6/data/html/generated/alphabet/index.html>

Automatické zpracování jazyka

- Programy, založené na **seznamech a pravidlech**
 - Spousta ruční práce počítačových lingvistů
 - Pravidla nikdy nepokryjí vše
- Programy, založené na **statistických modelech**
 - Velká „trénovací“ data
 - Velké nároky na výkon počítače

Od znaků po morfémy

- Tokenizátor (hranice slov)
- Segmentace na věty (příklad problému: „sv. Jan“)
- Morfologický analyzátor (např. Hajič, Majka)
 - slovník →
 - lemmata
 - morfologické značky
- zatím nejednoznačně, bez ohledu na kontext

Tagger (značkovář)

- Zjednotná morfológickou analýzu na základě kontextu
- Statistický model na ručně anotovaných trénovacích datech
- Pro češtinu např. Hajič HMM, Morče, **Featurama**
- Úspěšnost přes **95 %** správně označovaných slov

Parser (syntaktický analyzátor)

- Vstup: zjednoznačená morf. analýza z taggeru
- Výstup: strom (pro češtinu většinou **analytický** dle PDT)
- Statistický model na ručně anotovaných trénovacích datech (PDT)
- Pro češtinu např. **MST Parser**, **Malt Parser**
- Úspěšnost asi **86 %** správně zavěšených slov
 - Pokud započítáme i správnost afunů, klesne to k 80 %

Treex

- Použití nástrojů náročné
 - na obsluhu
 - na výkon počítače
 - datové formáty, spolupráce nástrojů
- Treex je naše zastřešující platforma
- Díky **webovému rozhraní** ho může použít každý
<http://ufal-point.mff.cuni.cz/services/treex-web/>

Shrnutí

- Hajičova morfologie
- Morfo (generování tvarů)
- Interset (rozklad a převod morfologických značek)
- Treex (taggery, parsery, překlad a další)
- Tred (prohlížeč a editor stromů)
- PML Tree Query (vyhledávání v treebancích)
- Česílko
- Korektor
- Styx a Čapek...