

IB047

Morfologické a syntaktické
značkování korpusů

Pavel Rychlý

pary@fi.muni.cz

24. Březen 2025

Pavel Rychlý IB047

Universal Dependencies

- nástupce Universal Tagset
- verze 2.13 (11/2023): 259 korpusů, 148 jazyků
- 14 značek – slovní druhy (+ 3 ostatní)
- open class
ADJ: adjective, ADV: adverb, INTJ: interjection, NOUN: noun, PROPN: proper noun, VERB: verb
- closed class
ADP: adposition, AUX: auxiliary, CCONJ: coordinating conjunction, DET: determiner, NUM: numeral, PART: particle, PRON: pronoun, SCONJ: subordinating conjunction
- ostatní
PUNCT: punctuation, SYM: symbol, X: other
- 21 features
PronType, Gender, Animacy, Number, Tense, Abbr, ...

Pavel Rychlý IB047

Morfologické značkování češtiny

Brněnské značky

- značka tvořena sadou znakových párů: atribut-hodnota
- atribut malým písmenem (k,g,n,c,p,x,w)
- hodnota velkým písmenem nebo číslicí
- k1 - k9, k0 – standardní slovní druhy ze ZŠ/SŠ
- kl – interpunkce
- k1gInSc4 = podstatné jméno, rod mužský neživotný, jednotné číslo, 4. pád
- korpus Desam

Pavel Rychlý IB047

Morfologické značkování

- každý token – značka
- několik desítek až tisíc značek (obsahující gramatické kategorie)
- Universal Tagset (Google)
12 značek – pouze slovní druhy
- jeden sloupec ve vertikálním tvaru

Pavel Rychlý IB047

Universal Dependencies

```
# newpar id = vesm9211-001-p7
# sent_id = vesm9211-001-p7s1
# text = Všechny tři světy si vzájemně trvale povídají a ovliv:
# orig_file_sentence vesm9211_001#8
Všechny DET Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|
tři NUM Case=Nom|Number=Plur|NumForm=Word|NumType=Card
světy NOUN Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|
si PRON Case=Dat|PronType=Prs|Reflex=Yes|Variant=Short
vzájemně ADV Degree=Pos|Polarity=Pos
trvale ADV Degree=Pos|Polarity=Pos
povídají VERB Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Polar:
a CCONJ -
ovlivňují VERB Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Polar:
se PRON Case=Acc|PronType=Prs|Reflex=Yes|Variant=Short
```

Pavel Rychlý IB047

Morfologické značkování češtiny

Brněnské značky

Z	z	k7c2
téměř	téměř	k6xMd1
tři	tři	k4xCgFnPc2
desítek	desítka	k4xNgFnPc2
smluv	smlouva	k1gFnPc2
upravujících	upravující	k2gFnPc2d1
vztahy	vztah	k1gInPc4
mezi	mezi	k7c7
oběma	dva	k4xCgInPc7
subjekty	subjekt	k1gInPc7
celního	celní	k2gMnSc2d1
soustátí	soustátí	k1gNnSc2
jsou	být	k5mItPp3nPaI
okamžité	okamžitě	k6xMd1
vypověditelné	vypověditelný	k2gFnPc1d1
všechny	všechn	k3xUgFnPc1
.	.	kI

Pavel Rychlý IB047

Morfologické značkování češtiny

Pražské značky

- různé formáty používané na ÚFAL MFF UK
- zejména v Prague Dependency Treebank (PDT)
- základní formát: poziční – 15 atributů dle pořadí

Pavel Rychlý IB047

Morfologické značkování češtiny

Pražské značky

```
Všechny  všechen  PLIP1-----  
tři      tři      ClXP1-----  
světy    svět     NNIP1-----A----  
si       se       P7-X3-----  
vzájemně  vzájemně  Dg-----1A----  
trvale   trvale   Dg-----1A----  
povídají povídat  VB-P---3P-AA---  
a        a        J^-----  
ovlivňují  ovlivňovat  VB-P---3P-AA---  
se       se       P7-X4-----  
.        .        Z:-----
```

Pavel Rychlý IB047

Morfologický analyzátor majka

- morfologická databáze slov
- u každého slova: základní tvar (lemma), značka
- různé funkce – různé datové formáty
- generováno ze slovníku a popisu vzorů
- použití konečného automatu
- <https://nlp.fi.muni.cz/ma/>

Pavel Rychlý IB047

Morfologický analyzátor majka

příklad pro analýzu: vstup=*s/ovo*, výstup=lemma+značka

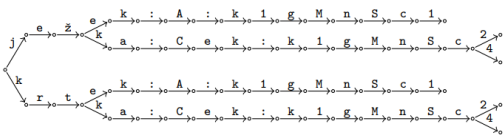
```
ježek:A:k1gMnSc1 <-- ježek:ježek:k1gMnSc1  
ježka:Cek:k1gMnSc2 <-- ježka:ježek:k1gMnSc2  
ježka:Cek:k1gMnSc4 <-- ježka:ježek:k1gMnSc4  
krtek:A:k1gMnSc1 <-- krtek:krtek:k1gMnSc1  
krtka:Cek:k1gMnSc2 <-- krtka:krtek:k1gMnSc2  
krtka:Cek:k1gMnSc4 <-- krtka:krtek:k1gMnSc4
```

Pavel Rychlý IB047

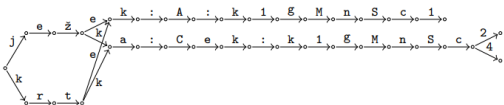
Morfologický analyzátor majka

Velká data převedena do konečného automatu

- non-minimalized deterministic automaton for the example data



- minimized deterministic automaton for the same data



Pavel Rychlý IB047

Morfologický analyzátor majka

Statistické informace

dictionary	lines	source MB	dictionary MB	bytes/line
w	13,609,590	186	3.3	0.240
w → l	14,101,767	240	4.0	0.287
w → l+t	80,303,929	2,478	4.4	0.054
w → w	957,464,060	19,993	6.1	0.006

Pavel Rychlý IB047

Použití pro slovníky

- základní tvar = heslo ve slovníku
- slovní druh
- další morfologické tvary slova
- časté použití (např. většinou v množném čísle)

Pavel Rychlý IB047

Základní tvar – lemma

- podstatná jména – 1. pád, jednotné číslo
- přídavná jména – 1. pád, jed. č., mužský rod
- slovesa – infinitiv
- ne vždy jednoznačné: milión/milion, myslit/myslet
- různé preference: nejlepší/dobry, neučit/učit
při vyhledávání na webu chceme něco jiného než lingvisté
- korpus.cz: řešení pomocí dvou atributů: lemma, sublemma

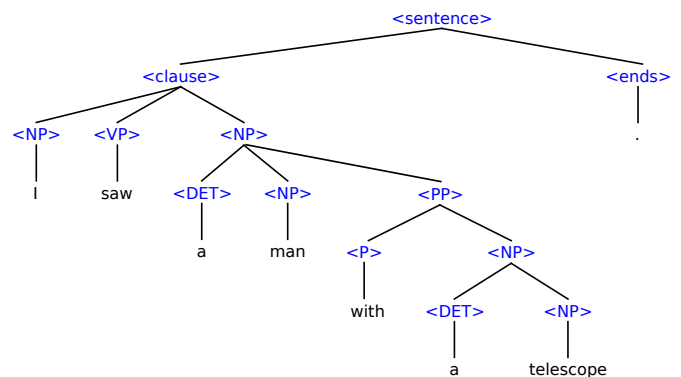
Pavel Rychlý IB047

Přístupy k syntaxi

- pro každou větu vytvoříme strom zachycující vztahy mezi slovy a/nebo skupinami slov
- frázový (složkový) postupně ze slov vytváříme skupiny
- závislostní určíme závislosti mezi jednotlivými slovy
- řádově náročnější problém: dokumentace často na stovkách stran

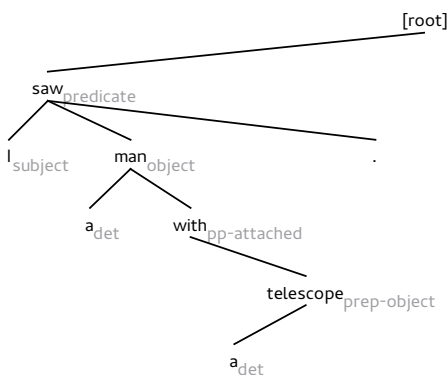
Pavel Rychlý IB047

Phrase structure formalism – example



Pavel Rychlý IB047

Dependency formalism – example



Pavel Rychlý IB047

Dependency vs. phrase-structure

Non-projectivity

- disconnected phrases
- not natural in the phrase structure notation
- 20% of Czech sentences are reported to contain a non-projective dependency

Phrase structure – more fine-grained analysis

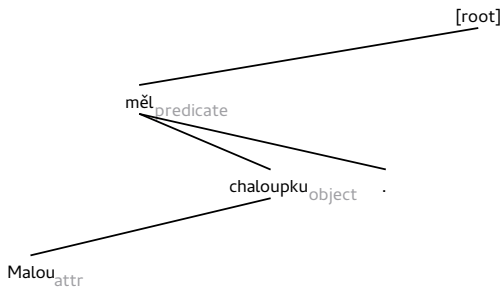
- (new (queen of beauty))
- (new generation)(of fighters)

Coordinations and other “flat” phenomena

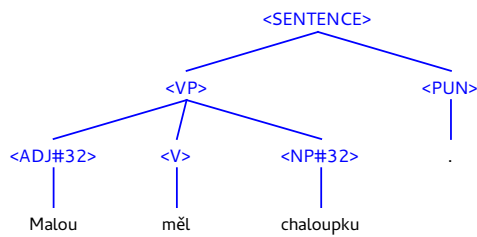
- not natural in the dependency notation
- problem for dependency analysis

Pavel Rychlý IB047

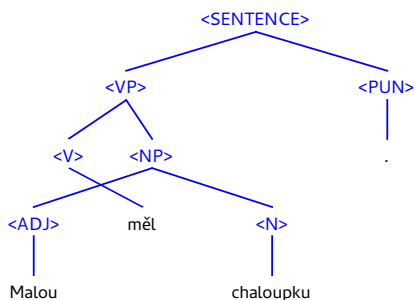
Non-projectivity – example



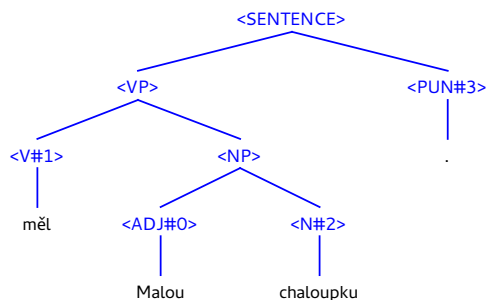
Non-projectivity in phrase structure formalism



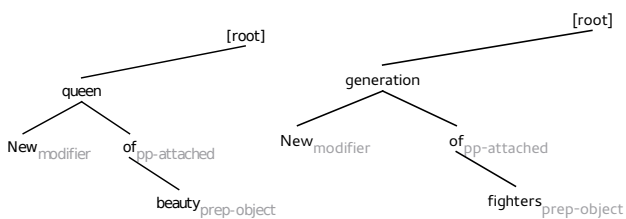
Non-projectivity in phrase structure formalism



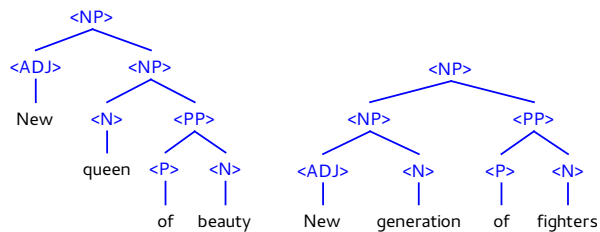
Non-projectivity in phrase structure formalism



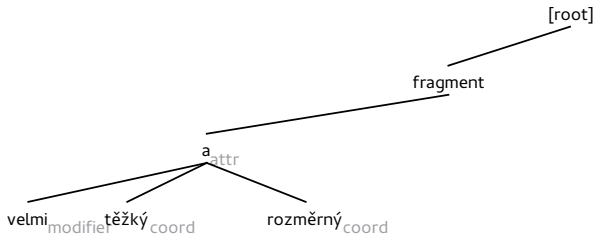
Phrase structure expressivity



Phrase structure expressivity

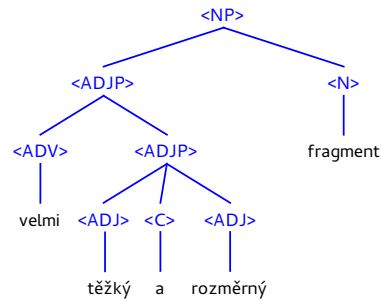


Coordinations – dependency structure



Pavel Rychlý IB047

Coordinations – phrase structure



Pavel Rychlý IB047

Záznam syntaxe v korpusu

- složkový systém:
 - fráze jako struktury – <phr>, </phr>
 - typy jako atributy
- závislostní systém:
 - očíslování tokenů, odkazy na tokeny
 - typy relací v dalším atributu

Pavel Rychlý IB047

Záznam syntaxe v korpusu

Složkový systém: Penn Treebank

```

(S (S (NP *)
  (VP Show
    (NP me)
    (NP (NP (NP all)
      the nonstop flights
      (PP (PP from
        (NP Denver))
        (PP to
          (NP San Francisco))))))
  (S (NP *)
    (VP leaving
      (NP about
        three
        o'clock
        (PP in
          (NP the afternoon))))))
  .)
  
```

Pavel Rychlý IB047

Záznam syntaxe v korpusu

Složkový systém: Susanne

A01:0260.27	NN1c	Implementation	[O[S[Ns:S.
A01:0260.30	IO	of	[Po.
A01:0260.33	NP1p	Georgia	[Ns[G[Nns.Nns]
A01:0260.39	GG	+<apos>s	.G]
A01:0270.03	NN1c	automobile	[Ns.
A01:0270.06	NN1n	title	.Ns]
A01:0270.09	NN1n	law	.Ns]Po]Ns:S]
A01:0270.12	YG	-	[c121.c121]
A01:0270.15	VBDZ	was	[Vsp.
A01:0270.18	RR	also	[R:G121.R:G121]
A01:0270.21	VVnt	recommend	.Vsp]
A01:0270.24	Ib	by	[Pb:a.
A01:0270.27	AT	the	[Ns.
A01:0270.30	JJ	outgoing	.
A01:0270.33	NN1c	jury	.Ns]Pb:a]S]
A01:0270.39	YF	+	.O]

Pavel Rychlý IB047

Záznam syntaxe v korpusu

Závislostní systém: PDT (Universal Dependencies)

#	orig_file_sentence	vesm9211_001#8			
1	Všechny	DET	3	det	
2	tři	NUM	3	nummod	
3	světy	NOUN	7	nsubj	
4	si	PRON	7	obl:arg	
5	vzájemně	ADV	7	advmod	
6	trvale	ADV	7	advmod	
7	povídají	VERB	0	root	
8	a	CCONJ	9	cc	
9	ovlivňují	VERB	7	conj	
10	se	PRON	9	obj	
11	.	PUNCT	7	punct	

Pavel Rychlý IB047