CJBB85 Počítačové nástroje pro češtinu

Syntaktická analýza

Jakub Machura

Masarykova univerzita
Filozofická fakulta
machura@phil.muni.cz

jaro 2021

NLP = Natural Language Processing

NLP = Natural Language Processing

Co je NLP?

NLP = Natural Language Processing

Co je NLP?

mluvené slovo × strojově čitelný text

NLP = Natural Language Processing

Co je NLP?

mluvené slovo × strojově čitelný text

analýza × syntéza jazyka

NLP

Kam zapadá syntax a syntaktická analýza?

Dílčí úkoly analýzy jazyka

Dílčí úkoly analýzy jazyka

Tokenizace

Dílčí úkoly analýzy jazyka

Tokenizace

„*Chcete-li mi to dát, neváhejte!*“

Tokenizace

„Chcete-li mi to dát, neváhejte!“

„
Chcete
-
li
mi
to
dát
,
neváhejte
!
“

Tokenizace

ohlas

Tokenizace

ohlas

- imperativ slovesa *ohlásit*
- nom./akuz. substantiva *ohlas*
- 2. os. sg. fem. minulého času slovesa *ohnout*

Větná segmentace

Větná segmentace

- explicitně vyznačený začátek i konec věty

Větná segmentace

- explicitně vyznačený začátek i konec věty

např. XML: <s> </s>

Větná segmentace

- explicitně vyznačený začátek i konec věty

např. XML: <s> </s>

Větná segmentace

Caesar byl zavražděn r. 43 př. Kr. Řím byl tehdy na pokraji převratu.

Větná segmentace

Caesar byl zavražděn r. 43 př. Kr. Řím byl tehdy na pokraji převratu.

Větná segmentace

Caesar byl zavražděn r. 43 př. Kr. Řím byl tehdy na pokraji převratu.

Jak to vyřešit?

Větná segmentace

Caesar byl zavražděn r. 43 př. Kr. Řím byl tehdy na pokraji převratu.

Jak to vyřešit?

Další problémy?

Morfologická analýza

Morfologická analýza

lemma, lemmatizace

Morfologická analýza

lemma, lemmatizace

Na	na
vyzvání	vyzvání (~~vyzvánět~~)
svého	svůj
předsedy	předseda
jsme	být
odešli	odejít (~~odeslat~~)
.	.

Morfologická analýza

lemma, lemmatizace

tag, tagging

Morfologická analýza

lemma, lemmatizace

tag, tagging, tagger

desambiguace

Tagging

- většinou založena na statistických modelech, někdy kombinováno s pravidly

Tagging

- většinou založena na statistických modelech, někdy kombinováno s pravidly
- ruční anotace

Tagging

- většinou založena na statistických modelech, někdy kombinováno s pravidly
- ruční anotace
- statistika četnosti značek

Tagging

- většinou založena na statistických modelech, někdy kombinováno s pravidly
- ruční anotace
- statistika četnosti značek
- „natrénování“ taggeru

Desambiguace

stochastická/statistická

Desambiguace

stochastická/statistická

založená na ling. pravidlech

Desambiguace

stochastická/statistická

založená na ling. pravidlech

hybridní

Desambiguace

Syntaktická desambiguace

Desambiguace

Syntaktická desambiguace

František hrál v altánu šachy se svým ruským přítelem.

Desambiguace

Syntaktická desambiguace

František hrál v altánu šachy se svým ruským přítelem.

Desambiguace

Sémantická desambiguace

Desambiguace

Sémantická desambiguace

využívat zařízení

Desambiguace

Sémantická desambiguace

využívat zařízení

dělat chyby ve skloňování

Parsing = Syntaktická analýza

Parsing

Cíle:

- „porozumět“ gramatice př. jaz.
- odhalit povrchovou strukturu (větný rozbor)

Parsing

Výsledky:

- orientované grafy (tzv. stromy)

závislostní × složkový

Parsing

Překážky:

- pro čj bohatá morfologie a rel. volný slovosled

Parsing

Překážky:

- pro čj bohatá morfologie a rel. volný slovosled
- **velké množství teoretických východisek**

Parsing

Překážky:

- pro čj bohatá morfologie a rel. volný slovosled
- velké množství teoretických východisek
- **subjektivita syntaxe**

Parsing

Překážky:

- pro čj bohatá morfologie a rel. volný slovosled
- velké množství teoretických východisek
- **subjektivita syntaxe**

Faxu škodí především přetížené telefonní linky.

Parsing

Víceznačnost:

Parsing

Víceznačnost:

1. Předložkové fráze (PP)

Parsing

Víceznačnost:

1. Předložkové fráze (PP)

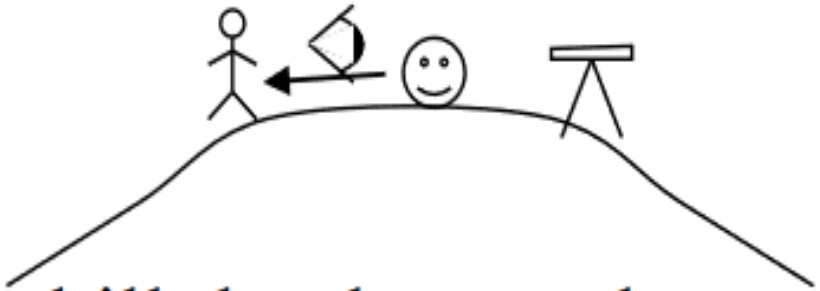
Charles talked about cooking with Britney Spears.

1. Předložkové fráze (PP)

I saw the man on the hill with the telescope.

1. Předložkové fráze (PP)

I saw the man on the hill with the telescope.



"I was on the hill that has a telescope when I saw a man."

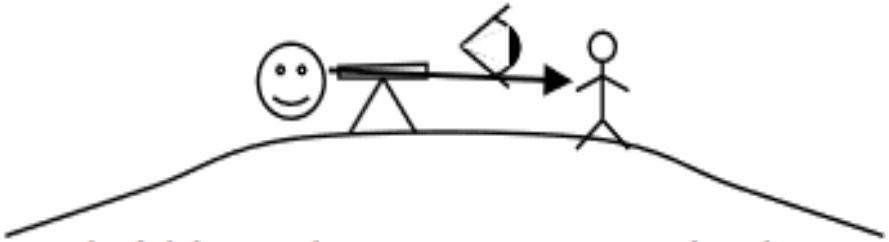
1. Předložkové fráze (PP)

I saw the man on the hill with the telescope.

"I saw a man who was on the hill that has a telescope on it."

1. Předložkové fráze (PP)

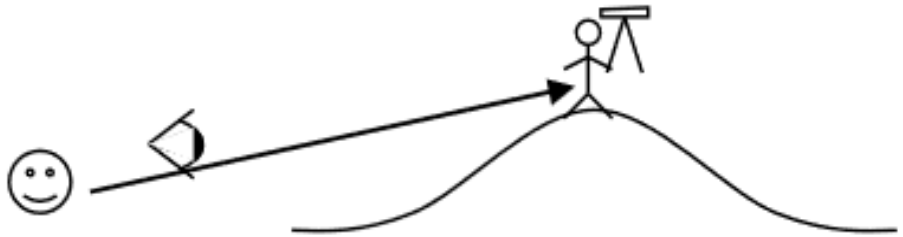
I saw the man on the hill with the telescope.



"I was on the hill when I used the telescope to see a man."

1. Předložkové fráze (PP)

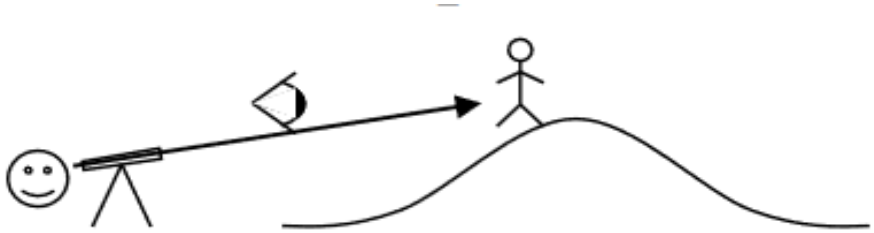
I saw the man on the hill with the telescope.



"I saw a man who was on a hill and who had a telescope."

1. Předložkové fráze (PP)

I saw the man on the hill with the telescope.



"Using a telescope, I saw a man who was on a hill."

2. Elipsa (gap)

Marie má ráda fyziku, ale nesnáší chemii.

[Mary likes Physics but hates Chemistry.]

2. Elipsa (gap)

Marie má ráda fyziku, ale nesnáší chemii.

[Mary likes Physics but hates Chemistry.]

3. Koordinační konstrukce

Small boys and girls are playing.

Dřevěná vrata a okna natřel nabílo.

4. Slovnědruhová homonymie

She ran up a large bill.

She ran up a large hill.

4. Slovnědruhová homonymie

She ran up a large bill. [částici]

She ran up a large hill. [předložka]

4. Slovnědruhová homonymie

She ran up a large bill. [částici]

She ran up a large hill. [předložka]

Umyl se úplně celý.

Umyl se žínkou nádobí.

4. Slovnědruhová homonymie

She ran up a large bill. [částici]

She ran up a large hill. [předložka]

Umyl se úplně celý. [zvratné zájmeno]

Umyl se žínkou nádobí. [předložka]

4. Slovnědruhová homonymie

Frightening kids can cause troubles.

[gerundium vs. adjektivum]

4. Slovnědruhová homonymie

Frightening kids can cause troubles.

[gerundium vs. adjektivum]

Zdraví nemocnému nevěří.

Zdraví si musíme chránit.

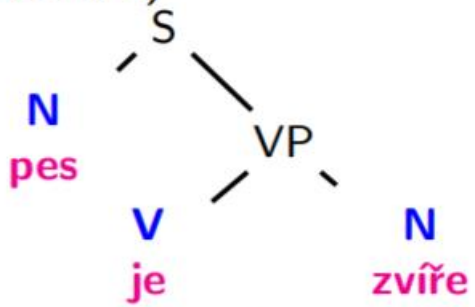
Zdraví vás z Krušných hor.

Základní termíny

větná struktura

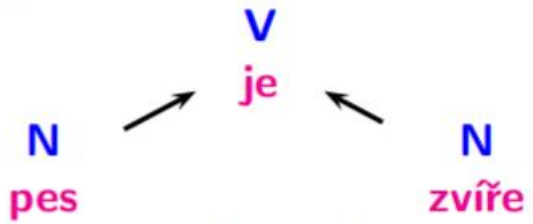
- povrchová struktura (*surface structure*)

  derivační/složkový strom jako výsledek bezkontextové (CF) analýzy



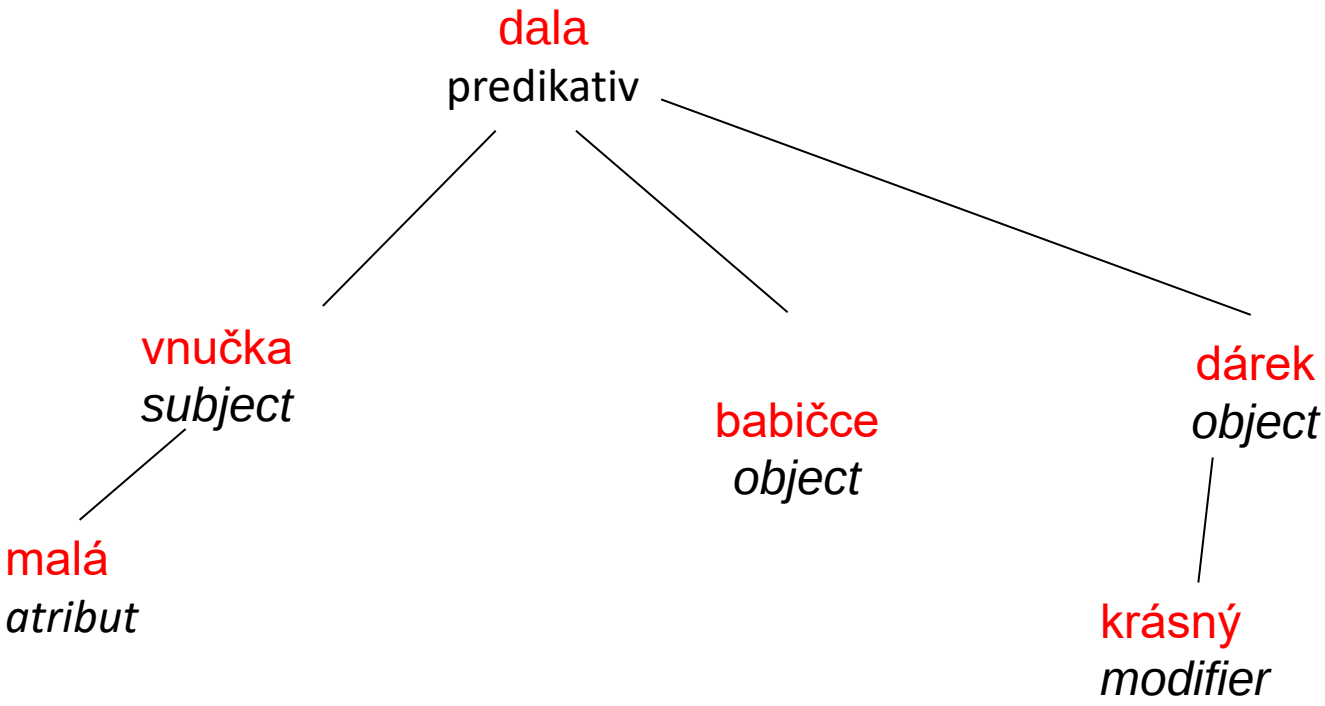
- závislostní struktura (*dependency structure*)

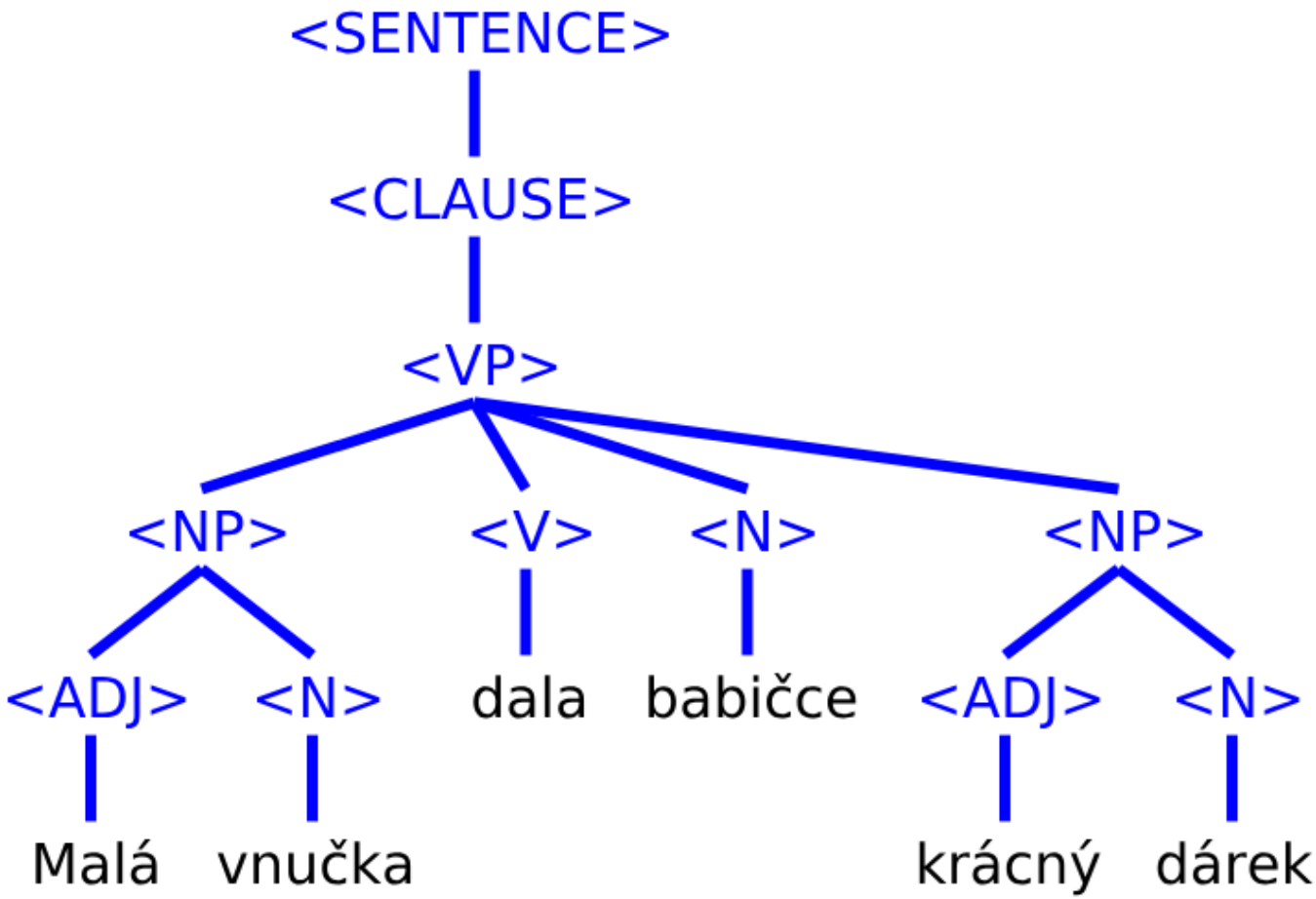
  zobrazuje závislosti mezi větnými členy



- hloubková struktura (*deep structure*) – sémantická interpretace fráze. Popisuje role větných členů (agens, patiens, donor, cause, . . .)

Malá vnučka dala babičce krásný dárek.




<SENTENCE>
<CLAUSE>
<VP>
<NP>
<V>
<N>
<NP>
<ADJ>
<N>
dala
babičce
<ADJ>
<N>
Malá
vnučka
krácný
dárek

Syntaktická analýza – Praha

- historické pozadí
- lingvistický strukturalismus, Pražská škola
- Pražský lingvistický kroužek (1926, Mathesius, Jakobson, Trnka)
- funkčně generativní popis (FGD, Petr Sgall, 60. léta)
  - závislostní syntax
  - hloubková (tektogramatická) struktura
  - formální popis aktuálního členění věty

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

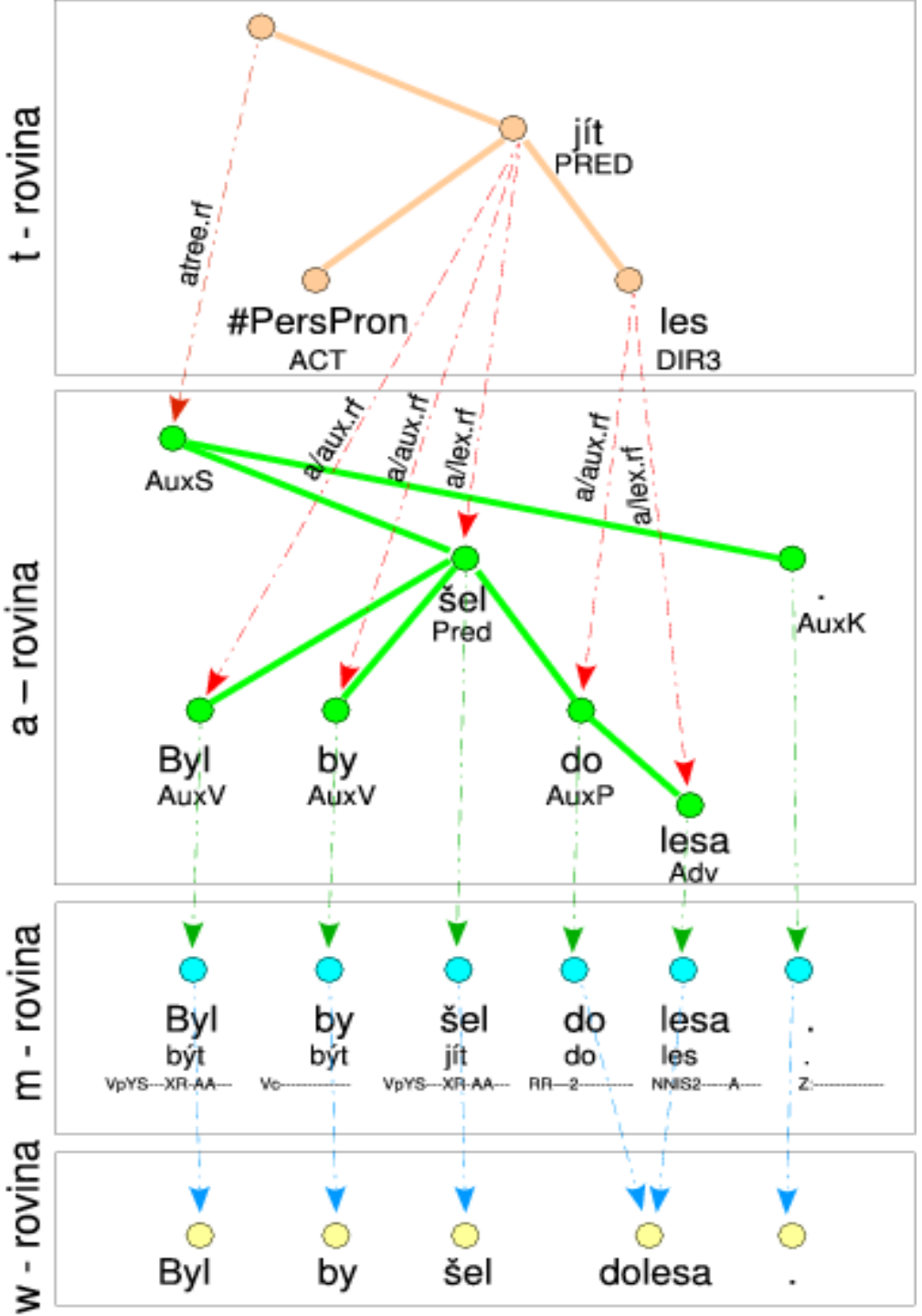
- morfologická rov.
  - 2 mil. slovních výskytů
- analytická rov.
  - 1,5 mil. slovních tvarů

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


t - rovina
atree.rf
jít
PRED
#PersPron
ACT
les
DIR3
a – rovina
a/aux.rf
a/aux.rf
a/lex.rf
a/aux.rf
a/lex.rf
AuxS
šel
Pred
.
AuxK
Byl
AuxV
by
AuxV
do
AuxP
lesa
Adv
m - rovina
Byl
být
VpYS---XR-AA---
by
být
Vc----------
šel
jít
VpYS---XR-AA---
do
do
RR---2----------
lesa
les
NNIS2-----A---
.
.
Z:----------
w - rovina
Byl
by
šel
dolesa
.

ZOBRAZENÍ ZÁVISLOSTNÍCH STRUKTUR

Syntaktická analýza – Brno

- FI MU – CZPJ
- syntaktický analyzátor klara, synt, set
- morfologicky značkovaný korpus
- formální popis gramatiky

Odkazy

- `synt`

  http://nlp.fi.muni.cz/projekty/wwwsynt/
- `set`

  https://nlp.fi.muni.cz/projekty/set/wwwset.cgi/first_page