Morfologická analýza
Desambiguace

Počítačové nástroje pro češtinu
Jaro 2019

Marie Novotná

Proč to potřebujeme?

- morfologické značkování korpusů
  - součástí vnitřní anotace korpusu
  - zvýšená informační hodnota
- vyhledávání v korpusu
- předpoklad pro další stupně analýzy jazyka
  - syntaktická, sémantická
- předpoklad pro navazující aplikace
  - např. Word Sketch Engine, Morfio
- zapojení do dalších nástrojů pro práci s jazykem
  - kontrola pravopisu, překladače, slovníky, webové prohlížeče
- možnost adaptace pro jiné slovanské jazyky

Základní pojmy

- morfologická značka (tag, index)
  - kód přiřazený k jednotlivým tvarům slov; nese informaci o morfologických charakteristikách slov
- tagset
  - soubor používaných morfologických značek
- značkování (tagování, tagging, anotace, indexování)
  - automatické, poloautomatické, manuální
- morfologický analyzátor (morphological analyzer, tagger)
- desambiguace (disambiguace, disambiguation)
  - zjednoznačnění, výběr správné morfologické značky v závislosti na kontextu
- guesser

Značkovací systémy (tagsety)

- poziční systém (Hajič, Hlaváčová)
  - „pražský“, ČNK
  - každá kategorie je určena pozicí
  - 16 pozic (do SYN2005 to bylo 15)
- atributivní systém (Osolsobě)
  - „brněnský“
  - dvojice atribut–hodnota bez ohledu na pořadí
  - přehlednější, úspornější, snadno rozšiřitelný
- heterogenní systém
  - „bratislavský“
  - vynechává prázdné pozice
  - nejkratší, ale malá rozšiřitelnost
- BNC tagset
  - pevná množina hotových značek
  - AJC Comparatice adjective, AJ0 Adjective (general or positive), ...

atributivní systém

k1 - Substantivum

x	Speciální vzor
P	půl

g	Rod
M	Rod mužský životný
I	Rod mužský neživotný
N	Rod střední
F	Rod ženský
R	Rodina (příjmení)

n	Číslo
S	Číslo jednotné
P	Číslo množné
D	Duál
R	Hromadné ozn. čl. rodiny

c	Pád
1	Pád první
2	Pád druhý
3	Pád třetí
4	Pád čtvrtý
5	Pád pátý
6	Pád šestý
7	Pád sedmý

Tvary

Značka	Typ tvaru
zS	tvar s příklonným s

Poznámky

Značka	Poznámka
rD	deverbativum
hF	ženské posesivum
hM	mužské posesivum
hR	rodinné posesivum
tQ	vyjadřuje míru
tA	vyjadřuje zřetel
tL	vyjadřuje místo
tT	vyjadřuje čas
tC	vyjadřuje příčinu
tM	vyjadřuje způsob
tD	příslovce modální
tS	příslovce stavové
hT	zastupuje věc (co?)
hP	zastupuje osobu (kdo?)
xC	číslovka základní
xO	číslovka řadová
xR	číslovka druhová

k2 - Adjektivum

e	Negace
A	Afirmace
N	Negace

g	Rod
M	Rod mužský životný
I	Rod mužský neživotný
N	Rod střední
F	Rod ženský

n	Číslo
S	Číslo jednotné
P	Číslo množné
D	Duál

c	Pád
1	Pád první
2	Pád druhý
3	Pád třetí
4	Pád čtvrtý
5	Pád pátý
6	Pád šestý
7	Pád sedmý

d	Stupeň
1	Positiv
2	Komparativ
3	Superlativ

Styl

w	Stylistický příznak
A	archaismus
B	básnicky
C	pouze v korpusech
E	expresivně
H	hovorově
K	knižně
O	oblastně
R	řídčeji
Z	zastarale

k3 - Zájmeno

p	Osoba
1	První osoba
2	Druhá osoba
3	Třetí osoba
X	1., 2. nebo 3.

g	Rod
M	Rod mužský životný
I	Rod mužský neživotný
N	Rod střední
F	Rod ženský

n	Číslo
S	Číslo jednotné
P	Číslo množné
D	Duál

c	Pád
1	Pád první
2	Pád druhý
3	Pád třetí
4	Pád čtvrtý
5	Pád pátý
6	Pád šestý
7	Pád sedmý

k4 - Číslovka

g	Rod
M	Rod mužský životný
I	Rod mužský neživotný
N	Rod střední
F	Rod ženský

n	Číslo
S	Číslo jednotné
P	Číslo množné
D	Duál

c	Pád
1	Pád první
2	Pád druhý
3	Pád třetí
4	Pád čtvrtý
5	Pád pátý
6	Pád šestý
7	Pád sedmý

Poznámky-pokr.

Značka	Poznámka
yQ	tázací
yR	vztažné
yN	zápor
yI	neurčitost
yF	reflexivní
xD	ukazovací
xT	vymezovací
xP	osobní
xO	přivlastňovací
xC	spojka souřadicí
xS	spojka podřadicí
c1	předložka s 1. pádem
c2	předložka s 2. pádem
c3	předložka s 3. pádem
c4	předložka s 4. pádem
c6	předložka s 6. pádem
c7	předložka s 7. pádem
aP	dokonavé
aI	nedokonavé
aB	obouvidé
wH	hovorově

k5 - Sloveso

e	Negace
A	Afirmace
N	Negace

a	Vid
P	Perfektivum
I	Imperfektivum
B	Obouvidé

m	Typ (Mód)
F	Infinitiv
I	Indikativ
R	Imperativ
A	Příčestí činné (minulé)
N	Příčestí trpné
S	Přechodník přítomný
D	Přechodník minulý
B	Indikativ futura

p	Osoba
1	První osoba
2	Druhá osoba
3	Třetí osoba

g	Rod
M	Rod mužský životný
I	Rod mužský neživotný
N	Rod střední
F	Rod ženský

n	Číslo
S	Číslo jednotné
P	Číslo množné

k6 - Příslovce

e	Negace
A	Afirmace
N	Negace

d	Stupeň
1	Positiv
2	Komparativ
3	Superlativ

k7 - Předložka

k8 - Spojka

k9 - Částice

k0 - Citoslovce

kA - Zkratka

kY - by,aby,. . .

m	Vztah k sl. módu
C	kondicionál

p	Osoba
1	První osoba
2	Druhá osoba
3	Třetí osoba

n	Číslo
S	Číslo jednotné
P	Číslo množné

poziční systém

1.	slovní druh	NAPCVDRJTIXZFHM
2.	detailní určení slovního druhu	!*,.:;=?^}~@ *[a-zA-Z0-9]*
3.	jmenný rod	-FHIMNQTXYZ
4.	číslo	-DPSWX
5.	pád	-X1234567
6.	přivlastňovací rod	-FMXZ
7.	přivlastňovací číslo	-PS
8.	osoba	-123X
9.	čas	-FHPRX
10.	stupeň	-123
11.	negace	-AN
12.	aktivum/pasivum	-AP
13.	nepoužita	-
14.	nepoužita	-
15.	varianta (stylový příznak)	-123456789
16.	vid	-PIB

České morfologické analyzátory

- analyzátor MORČE (MORfologie ČEštiny)
  - Jan Raab (ÚFAL MFF UK)
  - včetně desambiguace (statistický model)
  - morfologický slovník MorfFlex, tagger MorphoDiTa
- analyzátor AJKA (Analyzátor JazyKA)
- analyzátor MAJKA (Morfologický Analyzátor JazyKA)
- MorphCon (http://morphcon.webnode.cz)
  - převodník českých morfologických systémů
  - Pořízka, Schäfer, Zeman (Olomouc, Bonn, Praha)

Morfologický analyzátor Ajka

- Radek Sedláček
- formální (algoritmický) popis morfologie (Klára Osolsobě)
- systém atribut – hodnota
- slovo = řetězec znaků ohraničený z obou stran mezerami
- využívá struktury trie
- segmentace slova

  KmZ – IS – T

  kmenový základ, intersegment, koncovka
- koncovkové množiny
- slovník kmenů
- slovník intersegmentů
- seznam vzorů
- příliš složitá a nerozšiřitelná

Morfologický analyzátor Majka

- Pavel Šmerk
- princip konečných automatů
- jednodušší, rychlejší
- důkaz, že pro češtinu není třeba specializovaných datových struktur nebo algoritmů
- díky Majce vznikl podobný analyzátor pro slovenštinu
- rozšířena o slovenštinu, polštinu, angličtinu
- doplnění diakritiky CzAccent

Průběh morfologické analýzy

- rozeznání neohebných slovních druhů
  - po rozeznání analýza skončí
- rozeznávání slova od začátku
  - záporka ne-
  - superlativní prefix nej-
- segmentace slova od konce
  - koncovka
  - intersegment
  - kmenový základ
  - přiřazení ke vzoru

nej-ne-oblíben-ějš-ími

Desambiguace

- odstranění homonymie
- manuální, statistická (94 %), pravidlová, hybridní
- některé tvary nelze desambiguovat – ani na základě kontextu nelze jednoznačně přiřadit správnou značku

CQL

- Corpus Query Language
- dotazovací jazyk pro práci s korpusy
- formát [atribut=„hodnota“]
  - atribut – word, lemma, tag
  - hodnota – samotný výraz nebo výraz specifikovaný regulárním výrazem

[tag=„N.FS5.*“]

Regulární výraz

- řetězec popisující množinu řetězců
- nerekurzivní popis
- vyhledávání v textu pomocí zástupných znaků (metaznaků)
  - . ? * +
  - {n} {m,n} {m,}
  - \d \D \w \W \s \S
  - [A-Z] [1-5] (p|P)

[tag=„N.(F|I)S5.*“]

Procvičování

- [tag="V..P...3F.N.*"]
- [tag="N.MS5.*"]
- [tag="k1gMnSc5"]
- kočku
- seděli
- a
- oběma

Děkuji za pozornost!

Odkazy

- Ajka: http://nlp.fi.muni.cz/projekty/wwwajka
- MorphCon: http://morphcon.webnode.cz