

# Úvod do korpusové lingvistiky 3

1

**AUTOMATICKÁ  
MORFOLOGICKÁ ANALÝZA  
(TOKENIZACE, LEMMATIZACE  
A TAGGING,  
DESAMBIGUACE)**

# Rozdíl mezi tokenem a slovem

2

- slovo
- slovoforma/slovní tvar/textové slovo
- grafická forma a jednotka parole
- slovo a interpunkce
- slovo a složený tvar
- slovo a spřežka

# Tokenizace

3

- V korpusové lingvistice automatický proces, který člení text složený z písmen, interpunkčních znamének a mezer na jednotlivé izolované [tokens](#), tj. na slovní tvary a interpunkční znaménka pro účely dalšího (obvykle počítačového) zpracování.
- Řetězce znaků definované znakové sady a definované oddělovače

# Příklad

*„Chcete-li mi to dát, neváhejte!“*

4

”  
Chcete  
-  
li  
mi  
to  
dát  
,  
neváhejte  
!  
“

# Velikost korpusu

5

## syn2015

Popis:

Synchronní reprezentativní korpus

Velikost:

120 748 715 pozic

Webová stránka:

<http://wiki.korpus.cz/doku.php/cnk:syn2015>

Štítky:

čeština

reprezentativní

synchronní

řada SYN

psaný

Dostupná metadata:

Atributy		Struktury	
word	1 751 599	<doc>	3 376
lc	1 426 255	<text>	114 492
lemma	777 011	<p>	2 805 065
lemma_lc	733 708	<s>	8 004 732
tag	3 529	<hi>	539 521
pos	12	<lb>	48 905
case	8		

# Pozice

6

- Tokeny/pozice – jednotky, v nichž se měří velikost korpusu.
- Podle počtu pozic se přibližně vypočítává počet textových slov (s opakováním).
- 120 748 715 pozic 100 milionů slovních tvarů
- word – různé slovní tvary
- lc – různé slovní tvary, ignoruje se velikost písma

# Lemma

7

- Textové slovo/systemové slovo
- Slovo jako jednotka textu
- Slovo jako lexikální jednotka
- Lemma – základní tvar – reprezentativní tvar
- Reprezentativní slovníková podoba, která je při automatickém zpracování jazyka v procesu [lemmatizace](#) přidělována každé formě v [korpusu](#).

# Lemmatizace

8

- Přiřazení ↑lemmatu jednomu slovnímu tvaru (příp. skupině slovních tvarů) v textu.
- Může být součástí morfologické analýzy
- slovnímu tvaru se přiřadí všechna jeho lemmata nezávisle na kontextu (pro homonymní tvar může být takových lemmat více)



# Příklad

*Jí je špatně. Já sním o Vánocích všechno.*

9

- jí jíst/ona
- je být/ono
- špatně špatně
- . .
- já já
- sním sníst/snít
- o o
- vánocích vánoce
- všechno všechno
- . .

# Disambiguace

10

- V případě víceznačného výsledku je součástí lemmatizace [↗disambiguace](#) (zjednoznačnění) slovních tvarů v textu, kdy se náležité lemma stanoví z nabídky, kterou poskytla morfologická analýza, na základě kontextu.
- Problém disambiguace mnohdy nelze vyřešit na základě jazykového kontextu.

# Lemmatizátor

11

- Počítačový program, který provádí automatickou lemmatizaci .
- Je buď samostatný, n. je modulem morfologické analýzy, či morfologického [↑taggeru](#) provádějícího morfologickou [↑disambiguaci](#) textu.

# Význam lemmatizace

12

- umožní uživateli pracovat nikoli jen se slovními tvary, nýbrž i s lemmaty jakožto reprezentanty příslušných lexémů a jejich paradigmat
- usnadňuje práci s korpusem
- umožňuje pracovat s korpusem na vyšší rovině abstrakce

# tag

13

- Morfologická značka (běžně nazývaná **tag**) je sumarizací gramatické informace o hledaném slovu (pozici) v konkrétním kontextu. Možné tagy pro každý token se přiřazují na základě morfologické analýzy, výsledný tag je pak pro každý token vybrán během následné desambiguace.

# Tagger

14

- Počítačový program provádějící morfologickou [↑disambiguaci](#) textu, která je součástí [↑anotace textu](#) (obvykle v [↑korpusu](#)). K dispozici je: (a) **stochastický tagger** (statistický), který provádí disambiguaci na základě [↑strojového učení](#) ([👉Hajič & Hladká, 1997](#); [👉Hajič & Hladká, 1998](#); [👉Brants, 2000](#); [👉Votrubec, 2005](#)), n. (b) **tagger založený na lingvistických pravidlech**, která buď vytváří lingvista ([👉Karlsson & Voutilainen ad. \(eds.\), 1995](#); [👉Tapanainen & Voutilainen, 1994](#); [👉Oliva & Hnátková ad., 2000](#); [👉Květoň, 2006](#); [👉Petkevič, 2006](#)), n. se automaticky vyvozují z textů ([👉Brill, 1992](#)), n. (c) **hybridní tagger**, který spojuje výhody přístupů (a) a (b) ([👉Hajič & Krbec ad., 2001](#); [👉Jelínek & Petkevič, 2011](#)). Na výsledky činnosti **t.** obvykle navazuje program na syntaktickou analýzu textu, [↑parser](#). Někdy je **t.** součástí parseru.

# Příklad pos tagů

*Jí je špatně. Já sním o Vánocích všechno.*

15

- jí jíst/V/ona/P
- je být/V/ono/P
- špatně špatně/D
- . ./Z
- já já/P
- sním sníst/V/snít/V
- o o/R
- vánocích vánoce/N
- všechno všechno/P
- . ./Z

# Většinou je vše v pořádku

16

vyrazit - říkala , že se obléká , ale že **jí/P je/V špatně/D** . Když stále nešla , volal jsem jí znovu .  
v deset v práci , další říkala sestře , že **jí/P je/V špatně/D** , že bere inzulin , ale nemá ho s sebou  
svědků bydlících v okolí . „ Řekl jsem , že **jí/P je/V špatně/D** ... “ Útok žirafy Americká vědkyně Katy Williams a její  
„ čuchněte si , slečno ! “ a že **jí/P je/V špatně/D** a kolena že má rozedraná tak hrůzně , že mi  
mě poprosila , abych to vzala za ni , protože **jí/P je/V špatně/D** . Udělala jsem jenom to , co po mně chtěla  
probudila před svítáním . Její první myšlenka byla , že **jí/P je/V špatně/D** , protože jí zprostřed břicha vycházel silný pocit nevolnosti .  
Volala mu na mobil , aby hned přijel , že **jí/P je/V špatně/D** a kdovíco ještě . Prostě ho ukecala . Šel nahoru  
ní ale alkohol necítily a sama jim řekla , že **jí/P je/V špatně/D** a je diabetička . Ženu se podařilo zachránit včas .  
; krmí se sama ; přejedením si způsobí , že **jí/P je/V špatně/D** ; má strach , že bude muset přerušit kojení dítěte  
se bála , že nastydne , si mysleli , že **jí/P je/V špatně/D** od srdce – navzdory tomu , že Naděžda Pavlovna trpěla  
snad pět let a když zrovna neřídí auto , tak **jí/P je/V špatně/D** i při pětiminutové jízdě . . Nešikovně a nezkušeně jsem  
nemá křičet . Po aktu žena nabulíkovala pachateli , že **jí/P je/V špatně/D** a aby otevřel okno . “ Dívka v nestřeženém okamžiku  
 , která uvařila kávu ženě , jež tvrdila , že **jí/P je/V špatně/D** , přišla o pět tisíc korun . Cesty mohou být



# Občas se vloudí chyby

17

- MF plus** určuje a třídí . A přestože zná snad všechny , **jí/on/P je/být/V jenom** málokdy . „ Mě už na to neužije ,
- Sestry sudičky** nemytého uprchlíka , jak se láduje zbytky z kuřete , **jí/on/P je/být/V přímo** z talíře v lednici . Dospěla nakonec k názoru
- Týden** , ve Znojmě mají problémy s pravidelnou konzumací ryb ( **jí/on/P je/být/V jen** 33,1 % lidí ) . V Plzni mají ryby

" **Já/já/P sním/sníst/V o/o/R vánocích/vánoce/N bílých** , kdy každé přání splnilo se nám

# *Místo vašich schůzek sis měl lépe vybírat*

18

trvání oprav bude příjezd do zámku zcela znemožněný autům . **Místo/místo/R** *oprav bude* označené a ohraničené zábranami . Vstup do areálu

Bučicích . V ulici Nová budou prováděny úpravy silnice . **Místo/místo/R** *oprav bude* důkladně označeno . Cyklostezku dokončí letos Zruč nad

# Výhody lemmatizace

19

- Všechny tvary určitého lexému vyhledáme i bez lemmatizace:
- [lc="pes|ps([auiyů]|em|ům|ech|ovi|ové)"]
- [lemm="pes"]
- Pokud budeme chtít hledat všechna maskulina životná, která končí v nominativu sg. na s a skloňují se podle typu *pán*, bez lemmatizace a taggování to nepůjde.

# Tagset

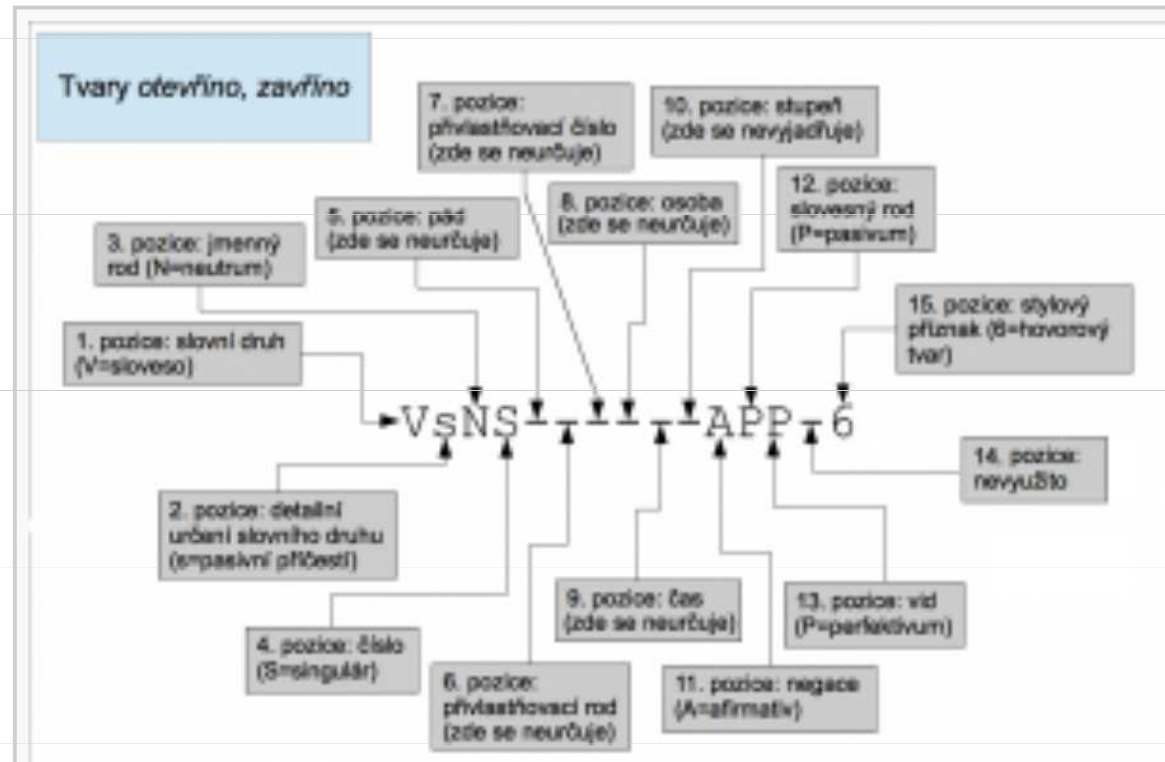
20

- Poziční
- atributový
- Gramatické vlastnosti, jako např. slovní druh, rod, číslo, pád, osoba, ..., se značkují a) na určité pozici nebo b) jsou pojmenovány určitým atributem.
- Gramatické významy jako např. substantivum, maskulinum, plurál, genitiv, 1. osoba, ..., jsou vyjádřeny závazně a) na určité pozici nebo b) za odpovídajícím atributem.

# poziční značka

21

## Struktura značky



Popis tagu odpovídajícího slovním tvarům *otevřino*,  
*zavřino* v korpusu [SYN2020](#)



## Různé korpusu a různé jazyky používají různé tagsety

- [https://wiki.korpus.cz/doku.php/cnk:intercorp:verz\\_e10#morfosyntakticka\\_annotace](https://wiki.korpus.cz/doku.php/cnk:intercorp:verz_e10#morfosyntakticka_annotace)
- <https://www.sketchengine.eu/German-rftagger-part-of-speech-tagset/>
- [http://utkl.ff.cuni.cz/%7Erosen/public/stts\\_guide.pdf](http://utkl.ff.cuni.cz/%7Erosen/public/stts_guide.pdf)

# Příklad otázek v testu

23

- Vyjmenuj jednotlivé kroky automatického zpracování textu, jejichž výsledkem je lemmatizovaný a morfologicky taggovaný text.
- Jak zjistíme velikost korpusu?
- Uveďte příklady, kdy jeden slovní tvar lze interpretovat vícero lemmaty.
- Uveďte příklady, kdy jeden slovní tvar lze interpretovat jediným lemmatem, ale má více významu gramatických kategorií pádu a čísla.