# Úvod do korpusové lingvistiky 3



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

# Rozdíl mezi tokenem a slovem



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

# Tokenizace



- V korpusové lingvistice automatický proces, který člení text složený z písmen, interpunkčních znamének a mezer na jednotlivé izolované ↗tokeny, 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!“*



„
Chcete
-
li
mi
to
dát
,
neváhejte
!
“

# Velikost korpusu



## 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



- 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



- Textové slovo/systémové 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



- 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.*



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

# Disambiguace



- 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



- Počítačový program, který provádí automatickou lemmatizaci .
- Je buď samostatný, n. je modulem morfologické analýzy, či morfologického ↗taggeru provádějícího morfologickou ↗disambiguaci textu.

# Význam lemmatizace



- 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



- 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



- 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.*



| | |
|---|---|
| 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



| | | |
|---|---|---|
| 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 necítili 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



| | | | | |
|---|---|---|---|---|
| ☐ | 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/snist/V o/o/R vánocích/vánoce/N bílých , kdy každé přání splnilo se nám .

# *Místo vašich schůzek sis měl lépe vybírat*



| | | |
|---|---|---|
| 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



- Všechny tvary určitého lexému vyhledáme i bez lemmatizace:
- [lc="pes|ps([auiyů]|em|ům|ech|ovi|ové)"]
- [lemma="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

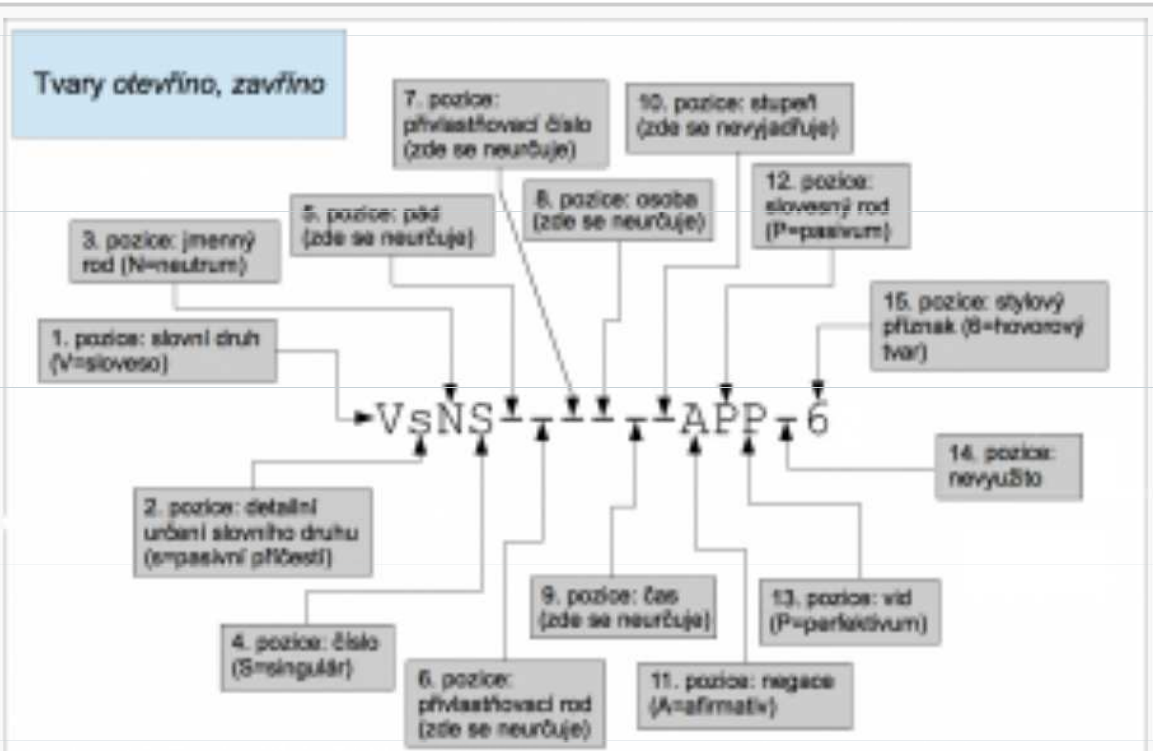


- Poziční atributový
- Gramatické vlastnosti, jako např. slovní druh, rod, číslo, pád, osoba, …, se značkují  a) na určité pozici nebo b) jsou pojmenovány určitým atributem.
- Gramatické významy jako např. substantivum, maskulinum, plurál, genitiv, 1. osoba, …, jsou vyjádřeny závazně a) na určité pozici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

# poziční značka



## Struktura značky




Popis tagu odpovídajícího slovním tvarům *otevřino, zavřino* v korpusu SYN2020

# Poziční systém tagu



Vlastní...

| | | | |
|---|---|---|---|
| á ještě neud | cesty do Kraslic . | Mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | , cítím se špatně . Nečekal jse |
| ním kouskem minulého roku nebo prvním nového . | | Mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | . Každé dopoledne čekám za ( |
| jsem se , že jsou ještě i levné požitky . | | Mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | . Jsem tu sám , skládám odve |
| zavolat a ujistit je , že jsem v pořádku a | | mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | . 12 . 15 : Jedna protivná serv |
| Vůbec nejsem naštvaná , " odpověděla naštvaně . " | | Mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | a jsem v pohodě . " No jistě , a |
| Wezr roztrhl obálku a četl : " Milá tetinko , | | mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | a posílám ty věci , už je nepot |
| jsi na tom líp než tvůj otec . " " | | Mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | , " řekl Wesley . " Když - no , |
| " Zaklonila hlavu , aby na něj viděla . " | | Mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | . A ty se máš ? " " Výborně . |
| . Zamumlá cosi jako , že se má dobře , | | mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | , protože mám přítelkyni , kter |
| Jonas . " Sijambo , " odpověděla Nikola . " | | Mám/mít/VB-S---1P-AAI-- se/se/P7--4--------- dobře/dobře/Dg-------1A---- | . " " Tak už mám dva žáky , " |

# Atributový systém tagu



| | | | | |
|---|---|---|---|---|
| ky na téma "jak se máš?</s><s>- děkuju, | **mám** mít/k5eAaImIp1nS | **se** se/k3xPyFc4 | **dobře** dobře/k6eAd1 | a jak se máš ty?</s><s>" byly vzhledem k |
| vičení.</s><s>Teď se to všechno obrátilo, | **mám** mít/k5eAaImIp1nS | **se** se/k3xPyFc4 | **dobře** dobře/k6eAd1 | a fitness mě živí," říká Filip Šteflovič o své |
| m, že jsem poslední... no neva Tak čauky, | **mám** mít/k5eAaImIp1nS | **se** se/k3xPyFc4 | **dobře** dobře/k6eAd1 | , jen se trochu už s těma staroušema nudír |
| dět na mé váze.</s><s>Je to sice pravda, | **mám** mít/k5eAaImIp1nS | **se** se/k3xPyFc4 | **dobře** dobře/k6eAd1 | , ale kila mě trápila stále.</s><s>Celý můj |
| ě rádi."</s><s>anketa,návštěvníci plesu: " | **Mám** mít/k5eAaImIp1nS | **se** se/k3xPyFc4 | **dobře** dobře/k6eAd1 | zpívám Karla Gota a Petra Muka, jinak nic. |
| Maminko NE!!!</s><s>7. týden:Maminko, | **mám** mít/k5eAaImIp1nS | **se** se/k3xPyFc4 | **dobře** dobře/k6eAd1 | .</s><s>Jsem v nebíčku... Strašně mi chyb |
| i nadpřirozeného řádu.</s><s>"Maminko, | **mám** mít/k5eAaImIp1nS | **se** se/k3xPyFc4 | **dobře** dobře/k6eAd1 | .</s><s>Jsem v nebíčku..." Otázka možno |

# Různé korpusu a různé jazyky používají různé tagsety



- https://wiki.korpus.cz/doku.php/cnk:intercorp:verze10#morfosyntakticka_anotace
- https://www.sketchengine.eu/German-rftagger-part-of-speech-tagset/
- http://utkl.ff.cuni.cz/%7Erosen/public/stts_guide.pdf

# Araneum Francogallicum Minus (French, 15.03) 120 M



| | | |
|---|---|---|
| ce ballet orchestré à Agadez puis Niamey . L' on | parle/parler/VER:pres | d' une demande d' amnistie individuelle émanant de notre front |
| laisser la place aux multinationales ? Au moment où on | parle/parler/VER:pres | de traduire les responsables Guinéens auteurs des crimes et viols |
| en France de 1875 , les alternatives dont nous allons | parler/parler/VER:infi | au cours de ce séminaire apparaissent dans le système pénal |
| de dyslexie souffrant de légers troubles de la vue ne | parlant/parler/VER:ppre | pas français couramment . A ce jour là , BrowseAloud |
| sur le mot " diaporama " ! : Observons et | parlons/parler/VER:pres | de tout ce qui concerne notre planète et son humanité |
| et se pose continuellement des questions : " De quoi | parle/parler/VER:pres | l' histoire ? Que va-t-il arriver maintenant ? A quoi |
| qui se battent dans le parking du Fuzzy . On | parle/parler/VER:pres | encore de la légendaire bataille de Lacolle . On sait |
| éritage et une affirmation tragique de son existence , Marcel | parle/parler/VER:pres | ! Il parle de son enfance , sans père , |
| affirmation tragique de son existence , Marcel parle ! Il | parle/parler/VER:pres | de son enfance , sans père , dans une famille |
| monologue absurde , naïf , ironique et poétique , Marcel | parle/parler/VER:pres | aussi de ses peurs imaginaires et des dangers réels en |

# Araneum Germanicum Minus (German, 15.02) 120 M



| | | |
|---|---|---|
| erforderlich Argumente und Ziele die für den Erhalt des Bahnhofs | sprechen/sprechen/VVINF | „ Hier Brilon-Wald , hier Brilon-Wald ! “ bekannter Bahn |
| Theorie selbst erfassen , mit denen man also über Theorien | spricht/sprechen/VVFIN | . Definition 2.6 . Metatheoretische Begriffe Begriffe mi |
| eoretische Begriffe Begriffe mit denen man über Eigenschaften von Theorien | spricht/sprechen/VVFIN | werden metatheoretische Begriffe genannt . 2.3 .3 . Ol |
| die Sprache , mit der eine Wissenschaft über ihre Gegenstände | spricht/sprechen/VVFIN | . Betrachten wir zunächst die Verwendung der Wissen |
| Chemiker , z.B. , verwendet Sprache um über Gegenstände zu | sprechen/sprechen/VVINF | , die keine Sprache sind . Die Sprache , die |
| verwendet . Er verwendet Sprache , um über Gegenstände zu | sprechen/sprechen/VVINF | , die selbst nicht Sprache sind . Die Sprache , |
| rache , mit der man über nicht-sprachliche Gegenstände einer Wissenschaft | spricht/sprechen/VVFIN | , wird Objektsprache genannt . Der Wissenschaftler ve |
| Wissenschaftler verwendet Sprache jedoch auch , um über Sprache zu | sprechen/sprechen/VVINF | , z.B. über die Objektsprache seiner Wissenschaft . Da |
| 2.8 . Metasprache Jede Sprache mit der über eine Sprache | gesprochen/sprechen/VVPP | wird , ist eine Metasprache Da eine Metasprache selbs |
| Man kann in der Metasprache über alle Beschaffenheiten ihrer Objektsprache | sprechen/sprechen/VVINF | . Man kann in der Metasprache Regeln für den Gebrau |

# Příklad otázek v testu



- 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.