# CJDSL001 Korpusová lingvistika (2)

Klára Osolsobě

osolsobe@phil.muni.cz

Experimentální a počítačová lingvistika

# O čem budeme mluvit v kurzu

- Krátký historický exkurz
- Definice korpusu v moderním slova smyslu
- Dva metodologické přístupy k vytěžování korpusu
- Dva pohledy na korpus (lingvista a informatik)
- Filologie a korpusy
- Výuka jazyků a korpusy

# Definice korpusu v moderním slova smyslu

- Elektronické uložení
- Elektronická přístupnost
- Definovaný obsah (ČEHO) a rozsah (KOLIK)
- Standardní anotace – metada a interpretace jazykových jednotek
- Rychlost, spolehlivost a opakovatelnost vyhledávání a kvantifikace nalezeného

# Dva pohledy na korpus (lingvista a informatik)

- Nástroje NLP a korpusy
- Konverzní programy, vertikál, tokenizér
- Automatické analyzátory
- Lingvistické interpretace v korpusech

# Lingvistické informace v korpusech

- word/lc (slovoforma, slovní tvar, textové slovo)
- lemma (základní tvar, systémové slovo, položka ve slovníku, heslové slovo)
- tag (slovní druh a slovnědruhově závislé gramatické kategorie)
- K čemu je lemmatizace a taggování?
- Jaké jsou způsoby zaznamenání/kódování gramatických informací?
- Jaké jsou problémy a jak s nimi lze pracovat?

# Tagset (instrukce pro kódování významů gramatických kategorií)

https://wiki.korpus.cz/doku.php/pojmy:tag?redirect=1#tagset

https://www.sketchengine.eu/tagset-reference-for-czech/#toggle-id-1

Poziční systém značek

Atributivní systém značek

NNMS1----------

k1gMnSc1

# Automatická morfologická analýza / tagging

Tokenizace

Nejednoznačné přiřazení lemmatu a tagu (na základě slovníku)

Desambiguace

# lemma a lemmatizace

Lemmatizace - přiřazení systémového slova / základního tvaru slova k textovému slovu / slovnímu tvaru / slovoformě

Čím je tvarosloví bohatší, tím je lemmatizace žádoucnější

Základní tvar / lemma odpovídá slovníkovému tvaru

Diskutabilní jsou různé případy: a) spřežky (jeden grafický útvar představuje více slovních tvarů, a tedy i potenciálně více lemmat); b) analytické tvary a tzv. jednoslovná morfologie; c) lemmatizace slov, která nemají standardní základní tvar (např. jaké je lemma tvarů *pozdě* ***bycha*** *honit, křížem* ***krážem***, ***panečku***, ...); d) nakolik koresponduje lemmatizace v korpusu a tradiční hnízdování / paradigmatické tvoření (např. syntetické stupňování, verbální substantiva, adjektivizovaná participia, přechylování, posesivní adjektiva, ...); e) lemmatizace z hlediska variant a variantní lemma

Pokud je zvoleno technické řešení na úkor lingvistické tradice/školy/teorie, pak by mělo být konzistentně dodrženo pro všechny analogické případy a vysvětleno na úrovni uživatelských příruček.

# Vyhledání slovního tvaru *jít*

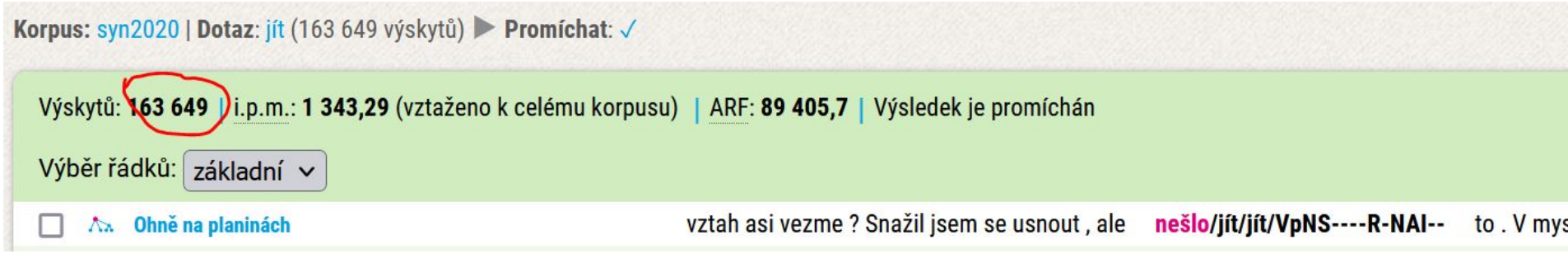
| | | |
|---|---|---|
| " Můj partner na rakovinu zemřel , protože se bál | jít | k lékaři . Z vlastní zkušenosti tedy vím , jak |
| , “ uvedl Malášek . Nevyloučil možnost , že může | jít | o vojáka nebo policistu . K druhé loupeži došlo jen |
| pádu . ^ * S výjimkou případů , kdy má | jít | k veterináři a vy se ji marně snažíte napěchovat do |
| mrazu . ( Lidovci už ztrácejí i Moravu . Zkusí | jít | do měst a zezelenat Jihomoravští lidovci získali v komunálních volbách |
| patriotům se změna názvu zřejmě líbit nebude , nemusí však | jít | o řešení trvalé . Vše nyní i v budoucnosti závisí |
| . Tyto potíže se léčí laserem . Po operaci může | jít | pacient hned domů Prof. MUDr. Pavel Kuchynka , CSc. , |
| v žádném případě . Je pravda , že jsme mohli | jít | na Švédsko a Kanadu . A nebudu lhát , že |
| v jedenáctimilionové zemi . ATÉNY „ Nevíme , kam máme | jít | , přijeli jsme lodí z Lesbosu , moje žena je |
| více Dominiků , než bývá obvyklé . A nemuselo hned | jít | o zapálené hokejové fanoušky . Stačí , že se jméno |
| unie přijala směrnici o nedovolené podpoře podnikání , která neumožňuje | jít | s nabídkou pod určitou minimální částku , " vysvětlil mluvčí |
| řadových odborářů , s jakým argumentačním mandátem k němu mají | jít | . Doslova v odborářském šoku jsem po té , co |
| trasy k nejzajímavějším místům parku . Můžete použít lodě , | jít | pěšky nebo jet džípem a bahnem se dobrodit až k |

# Vyhledání lemmatu *jít (KWIC+lemma+tag)*

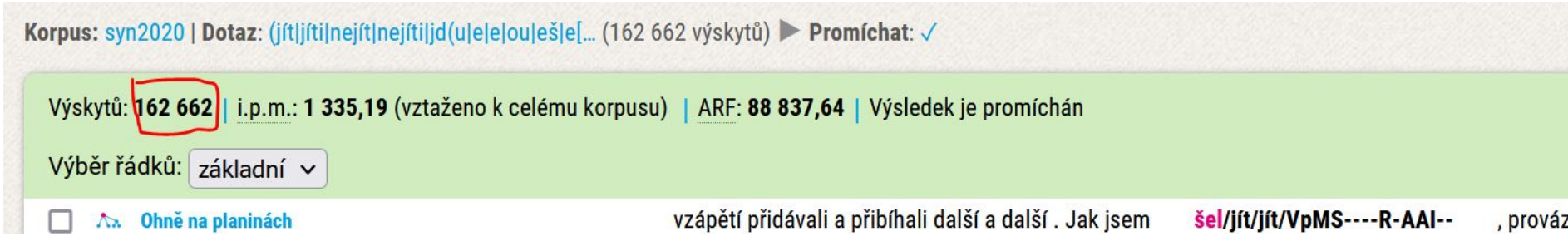
| | | |
|---|---|---|
| v Ledči . Pokud dva tři dny pořádně prší , | jde/jít/VB-S---3P-AA---I | začít pár kilometrů výše proti proudu a zdolat Stvořidla . |
| to přece dělal baron Prášil ! Americký kaskadér David Smith | jde/jít/VB-S---3P-AA---I | v jeho stopách - o víkendu tenhle bláznivý kousek předvedl |
| takové se mezi nimi možná občas najdou - , ale | jde/jít/VB-S---3P-AA---I | o spory , jež jdou nutně ruku v ruce s |
| ale také ceny za téměř třicet tisíc korun . " | Jde/jít/VB-S---3P-AA---I | opravdu o nálož cen kvalitních značek , " řekl pořadatel |
| připlácet , když chce státní úřad . Zvlášť , když | jde/jít/VB-S---3P-AA---I | o peníze , které tak jako tak od státu předtím |
| v baru , bowling . Něco , co lidi přiměje | jít/jít/Vf--------A----I | tam , a ne trávit čas jinak . Vesecko nabízí |
| s onou šaškárnou . Upevňuje se klamné přesvědčení , že | jde/jít/VB-S---3P-AA---I | o replikaci bytostí . Že je to něco jako nesmrtelnost |
| ání firemního mobilního telefonu a automobilu k soukromým účelům | jdou/jít/VB-P---3P-AA---I | ruku v ruce s rostoucími ekonomickými výsledky společnosti . Zkrátka |
| filmy promítají . „ Máme rozpis , podle kterého měla | jít/jít/Vf--------A----I | kopie filmu od nás do Sezimova Ústí . Bohužel jsem |
| byl opilý . Tehdy mi řekl , že když to | nepůjde/jít/VB-S---3F-NA---I | po dobrém , půjde to po zlém , “ vypověděla |
| na Liberec ? No , dlouho to vypadalo , že | jdeme/jít/VB-P---1P-AA---I | na Karlovy Vary , protože Litvínov v posledním kole vyrovnal |
| zjevně mladší a nezdravě agresívní . " Jsem ženská a | jdu/jít/VB-S---1P-AA---I | z práce . " odpověděla jsem poněkud mimo . " |
| v koupelně a přistihla je tam Kate , když si | šla/jít/VpFS---3R-AA---I | upravit nalíčení . Pak ale Pippa začala chodit s bankéřem |
| neexistující exoty a kdo všechno o tom podvůdku věděl a | nešel/jít/VpMS---3R-NA---I | to oznámit na policii . Ztratit paměť je , milí |
| , protože se jedná o klasické inflační peníze , které | jdou/jít/VB-P---3P-AA---I | v podstatě do stavebních firem . Mnohem lepší podle mne |
| = hrůza ! Jaké máte zkušenosti s českými silnicemi ? | Nejde/jít/VB-S---3P-NA---I | mi dost dobře do hlavy stav českých dálnic , které |
| v síti STEP ( ) . HLEdEJtE spoLEčnou cEstu Pokud | jde/jít/VB-S---3P-AA---I | o předkládání projektů na odstraňování starých ekologických zátěží , nejbližší |
| . Celý dospělý život jsem měla největší přání - aby | šel/jít/VpMS---3R-AA---I | bolševik od válu . A to se mi splnilo . |

# Lze vyhledat systémová slova bez pomoci lemmatizace / v nelemmatizovaném korpusu?

- [lemma="jít"]




- [lc="(jít|jíti|nejít|nejíti|jd(u|e|e|ou|eš|e[mt]e)|(ne|pů|nepů)jd(u|e|e|ou|eš|e[mt]e)|šel|šl[aoiy]|nešel|nešl[aoiy]|jd(i|ě[mt]e)|nejd(i|ě[mt]e)|nechoď|nechoď[mt]e|pojď|pojď[mt]e)"]



# Na co se zapomnělo?

| | | | | |
|---|---|---|---|---|
| 19 | p / n | nepůjde | 1440 | |
| 20 | p / n | šly | 1436 | |
| 21 | p / n | půjdou | 1260 | |
| 22 | p / n | nešel | 998 | |
| 23 | p / n | jděte | 982 | |
| 24 | p / n | jdete | 961 | |
| 25 | p / n | jdeš | 895 | |
| 26 | p / n | nešla | 780 | |
| 27 | p / n | půjdeš | 688 | |
| 28 | p / n | půjdete | 634 | |
| 29 | p / n | de | 590 | |
| 30 | p / n | nejdou | 495 | |
| 31 | p / n | nepůjdu | 317 | |
| 32 | p / n | nejdu | 285 | |
| 33 | p / n | nešli | 245 | |
| 34 | p / n | jdem | 245 | |
| 35 | p / n | nešly | 196 | |
| 36 | p / n | nepůjdeme | 173 | |
| 37 | p / n | půjdem | 171 | |
| 38 | p / n | nepůjdou | 149 | |
| 39 | p / n | di | 133 | |

# BOHUŽEL

**Korpus:** syn2020 | **Dotaz:** jít (163 649 výskytů) ▶ **Promíchat:** ✓ ▶ **Pozitivní filtr:** di (133 výskytů)

Výskytů: **133** | i.p.m.: **1,09** (vztaženo k celému korpusu) | ARF: **51,52** | Výsledek je promíchán 1 / 4

Výběr řádků: základní

| | | | | |
|---|---|---|---|---|
| Právo | se chodí koupat do římských fontán . Grillo a Luigi | Di/jít/jít/Vi-S---2--A-I-7 | Maio , jeden z hlavních činitelů hnutí a náměstek předsedy |
| A zničte Paříž! | celý vesnici , ty jedno pitomý jelito ! “ „ | Di/jít/jít/Vi-S---2--A-I-7 | do hajzlu , “ řekl jsem . „ Vypadni , |
| Blesk | hrají ale rozličné charaktery . Diváci se na nové díly | Di/jít/jít/Vi-S---2--A-I-7 | mohou těšit už příští rok . „ Mám pětadevadesát natáčecích |
| Duch | rádi pobývali Bill s Hillary a kam zavítala i Lady | Di/jít/jít/Vi-S---2--A-I-7 | a ze které dnes zbyly jen komíny , jenže tou |
| Zvěrolékař jde do boje | kladivem a dlátem , to je všechno . “ „ | Di/jít/jít/Vi-S---2--A-I-7 | ty , děláš si ze mě šoufky ! “ Simkin |
| Týden | , “ píše o Di Maiovi server The Local . | Di/jít/jít/Vi-S---2--A-I-7 | Maio má zjemňovat Grillovy výpady . To se ukázalo loni |
| Svět motorů Speciál | ) : Jako nový byl původní přímovstřikový diesel Endura - | Di/jít/jít/Vi-S---2--A-I-7 | terčem kritiky pro hrubý chod . Dnes se však ukazuje |
| Hospodářské noviny | milionů Italů , kteří nemají na jídlo , “ řekl | Di/jít/jít/Vi-S---2--A-I-7 | Maio . Někteří zákonodárci Hnutí pěti hvězd začínají mít obavy |
| Sport Góóól! | tisíc lidí . Pět procent zvolilo Klébersona , šest Ángela | Di/jít/jít/Vi-S---2--A-I-7 | Maríu , devatenáct Bebého a sedmdesát Alexise Sáncheze . Drtivý |
| Sport GÓÓÓL! | . Bude hrát za Fenerbahce v Turecku . Vzdal se | Di/jít/jít/Vi-S---2--A-I-7 | Marîi , 27letého Argentince a nejdražší posily v historii Red |
| Máslem dolů | . ! Zvládneš to , starý brachu ?! “ „ | Di/jít/jít/Vi-S---2--A-I-7 | do háje . “ říkám mu , ale to už |
| Nekropolis | Sráček probouzí , mně slečna Jessica řekla , José , | di/jít/jít/Vi-S---2--A-I-7 | se podívat na pánský záchodky , a tam to teda |
| Technický týdeník | Világi a zástupci dodavatelů – předseda představenstva Maire Tecnimont Fabrizio | Di/jít/jít/Vi-S---2--A-I-7 | Amato a generální ředitel Skupiny TALKE Alfred Talke . NOVÉ |
| Out | do očí , když s tebou mluvím ! “ „ | Di/jít/jít/Vi-S---2--A-I-7 | do prdele . Musíš začínat hned ráno ? “ „ |
| Game | a pokud ano , who benefited ? Jak zemřela princezna | Di/jít/jít/Vi-S---2--A-I-7 | , kdo se postaral , aby špion Litviněnko začal svítit |
| Pole | nosič a vší silou držet , aby mi neujel . | Di/jít/jít/Vi-S---2--A-I-7 | do prdele , řekne mi , ty zrádče ! Vo |
| Pole | si dobrej pozor , aby mě nezasáh . Ty sám | di/jít/jít/Vi-S---2--A-I-7 | do prdele ! řeknu já , stejně tvrdě jako on |
| Sport GÓÓÓL! | před rokem Van Gaal , 64letý Nizozemec , přivedl Ángela | Di/jít/jít/Vi-S---2--A-I-7 | Maríu ( 75 milionů eur ) , Lukea Shawa ( |
| Pole | to do prdele mluvíš ? zeptám se já . Prostě | di/jít/jít/Vi-S---2--A-I-7 | do prdele , buď tak hodnej ! Rozmáchne se , |
| Deníky Moravia | často zachovávají za cenu přesčasů . Přetížené sestry podle Veroniky | Di/jít/jít/Vi-S---2--A-I-7 | Cara z České asociace sester mohou při úkonech více chybovat |
| Sport GÓÓÓL! | aby můj tým zvítězil , “ říká . 03 Ángel | Di/jít/jít/Vi-S---2--A-I-7 | María V Anglii mu nemohli odpustit , že pořád leží |
| Zimní žár | svejch ze střílečky s Culpepperovejma , takže s takovejma řečma | di/jít/jít/Vi-S---2--A-I-7 | do háje . “ „ Hm , “ uvažoval Conner |
| Hospodářské noviny | spolupráce s Hnutím pěti hvězd . S jeho lídrem Luigim | Di/jít/jít/Vi-S---2--A-I-7 | Maiem se už začal domlouvat na tom , že si |

# Jak tedy funguje automatická morfologická analýza?

- Tokenizace (= rozdělení textu na jednotky, s nimiž bude nadále automatická analýza pracovat). Token se vždy nerovná textovému slovu. Textové slovo se vždy nerovná grafickému slovu (***Gazette de Villette***, ***Joffrey de Peyrac***, × ***Jak to de*?**).
- Přiřazení všech interpretací nalezených ve slovníku morfologického analyzátoru. NEJEDNOZNAČNOST výsledků dodaných ze slovníku.
- Desambiguace – výběr jediné interpretace (desideratum - kontextově odpovídající).

# Tagování v korpusu SYN2020
# https://wiki.korpus.cz/doku.php/cnk:syn2020:tag

| | | |
|---|---|---|
| zlomyslnost " ? ) a konkrétněji jeho skladba Il Pianto | di/di/di/F%---------- | Maria , zkomponovaná roku 1709 „ per il Venerdi Santo |
| Hradiště 3 . pprap . Bc . Michal Žáček - | DI/di/di/BN---------- | Kroměříž Pořadí 1 . Dopravní družstev : inspektorát Zlín 2 |
| penzionů nežijí žádní lidé . Nejstarší část Matery , Sassi | di/di/di/F%---------- | Matera ( Kameny Matery ) , je od roku 1993 |
| , rozsvítím lampu a dívám se na pohlednici z Madonny | di/di/di/F%---------- | Campiglio . Samozřejmě hned poznávám ledabylý rukopis svého syna . |
| str . 70 ) . Na český trh dodává produkty | di/di/di/F%---------- | - soric firma Amtek . Kapacitní snímače přiblížení di - |
| zda zasáhnou jeden z Potlouků nebo některého z protihráčů . | Di/di/di/F%---------- | - mitrov vyrazil přímo proti Moranové , která svírala Camrál |
| jsou v krátkých černých šatech z vlněné tkaniny od Giorgia | di/di/di/F%---------- | Sant Angela , pijí šampaňské s brusinkovým džusem a broskvovým |
| údajné podjatosti . Ukázky z knihy Z hrdiny nula h | di/di/di/F%---------- | I „ Znásilňovat holčičky se ve slušné společnosti prostě nenosí |
| " „ Co jsou Svobodní běžci ? " „ Ále | di/di/di/F%---------- | ! Vim , že ve Státech je máte taky . |
| a skokem ? To by šlo . " „ Už | di/jít/jít/Vi-S---2--A-I-7 | prosimtě , a hlavně zas nezakopni , ty chytráku . |
| jsem nezávodil od Gira . Následně se přesunu k Lago | di/di/di/F%---------- | Garda , kde se budu připravovat na další závody a |
| populistů může spustit snahy o vystoupení země z Unie . | Di/jít/jít/Vi-S---2--A-I-7 | Maio obavy mírní . „ Chceme zůstat v EU , |
| vzkaz , jen vstupenku do opery na Mozartovu Le Nozze | di/di/di/F%---------- | Figaro v Palais Garnier na dnešní večer . Zbrkle sáhla |
| Sráček probouzí , mně slečna Jessica řekla , José , | di/jít/jít/Vi-S---2--A-I-7 | se podívat na pánský záchodky , a tam to teda |
| . „ A - ano , do - dokud jsi | di/di/di/F%---------- | - divoch , " vypravil jsem ze sebe odpověď . |
| Jaroslava Plesla rovná se koncert citlivosti . Jarmilka a Bambini | di/di/di/F%---------- | Praga v interpretaci Oldřicha Kaisera rovnají se vy , Bohumile |
| blondýny s velkýma kozama v kovem zdobených šatech od Giorgia | di/di/di/F%---------- | Sant Angela . " Nepůjdem ji varovat ? " ptá |
| domem číslo 410 A v ulici Dumfries byl vzkaz : | DI/di/di/BN---------- | VEDLE Jakmile jsem přiběhl ke vchodu 410 B , nejdřív |
| tří stran obezděný výběžek ulice . Piazza de ' Cavalieri | di/di/di/F%---------- | Malta se náměstím téměř nedalo nazývat , spíš připomínalo nedostavěný |

# Proč je desambiguace obtížná?

- Pánové, *nežeňte se*
- Nemluv a *rožni*.
- Jan *je osel*.

# Víceznačné tvary

- *nežeňte/(ne)hnat/V*
- *nežeňte/(ne)ženit/V*
- *se/se/P*
- *se/s/R*
- *rožni/rožnit/V*
- *rožni/rozžehnout/V*
- *rožni/rožeň/N*
- *je/být/V*
- *je/on/P*
- *osel/osel/N*
- *osel/osít/V*

# hnát/Ženit

| | | |
|---|---|---|
| pracujte za byt a stravu . Ale nevdávejte se a | nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | jen kvůli tomu , že chcete uniknout z dor |
| sexu Shrnutí těchto vědeckých poznatků : Nejezděte na kole a | nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | s učitelkou . Potkáte - li ženu , která krá |
| legendární Knoflenkou . Takže se mějte a smějte ( a | nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | a nevdávejte ) . KDO S KÝM KDE |
| vážný důvod , proč tam chce jet . Tvůrci : | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | do ženění HEJNICE Pot |
| takovýchto " žertíků " neustále a iv jiném prostředí . | NEŽEŇTE/hnát/Vi-P---2--N----I SE/se/P7--4---------- | ZA SLEVOU , ŽÁDNÁ NENÍ Jedn |
| Uvědomte si , že vaše zásoba energie je omezená . | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | z jedné činnosti do druhé . 7 . V |
| Udo Pollmer říká : " Vyhoďte z bytu televizi a | nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | . Svatba je počátkem tukového obalovár |
| ventilaci - či pootevřete e dveře . Když skončíte , | nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | sprchovat - studenou vodou , ale zůstaň |
| to , co vyděláte hned dál investujte . 3 . | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | do nejistých investic Nezkoušín |
| požehnaného Čechova , jenž pravil , cituji : „ ' | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | , pokud se bojíte osamělosti . ' " <p: |
| se mohli vrátit k zaříkadlům , která přivolávají lásku . | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | však hned do vážných známostí , nejdřív |
| na cizí . Sebestian Roch Chamfort Bojíteli se samoty , | nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | ! Anton Pavlovič Čechov Bůh stv |
| Uvědomte si , že vaše zásoba energie je omezená . | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | z jedné činnosti do druhé . 7 . V |
| světě vidět a chovat vlastní dítě . Nepodléhejte trendům a | nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | na sál jen proto , že kamarádi byli . Zept |
| vědění je chromá . Věda bez náboženství je slepá . | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | za detaily . Nechť je vám v myšlenkách i |
| růst zajistí . 5 . Buďte trpěliví a nevzdávejte to | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | do velkých zakázek bezhlavě . Získávejt |
| 20.00 hodin kino v Hejnicích . Motto filmu zní : | Nežeňte/hnát/Vi-P---2--N----I se/se/P7--4---------- | do ženění . Až na svatební cestě Eddie z |

# být/on

| | | |
|---|---|---|
| Není to nic moc , ale Millicent to nepozná . | Je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | . Znova - que voulez-vous ? " Prozpěvovaly kvulevu , |
| Titulní strana Šípu , která o premiérovi hlásala , že | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | , či ta , kde trenér Brückner drží před začátkem |
| a profous považuje za svou povinnost říci mu , že | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | . Mužstvo jde na těžké cvičení do hor , ale |
| Nebudu dělat to , co řekne Manuel . „ Manuel | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | , Pere ! Osel je zvíře a ty přece nebudeš |
| mějský jarmark . V roce 1616 následovala hra Ďábel | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | , v níž ani ďábel nestačí na zlo lidí , |
| ocourskovský sedlák osít dvě míry pole krupicí . Kdo | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | , nechť se přihlásí . V říčce , do které |
| o rozesmálo všechny . Mě tedy nejvíc překvapilo , že | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | chytřejší než kůň . Po dotazech , které jsme měli |
| dy smím použít přejemnělého výrazu - jaký ten druhý | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | . Ale ne tak Evženka . Ta , když vypravuje |
| kde se před drátěným plotem něco pohybovalo . " To | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | ! " zajásala Julie a už se k němu rozeběhla |
| : A co dále , Baltazare , kde hlavním hrdinou | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | nebo u S. Paradžanova : Barva granátového jablka , kde |
| osla a pro jistotu pod obraz napsal : " Toto | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | . " Bedlivější analýzou vyprávěných anekdot lze zjistit , že |
| to na tebe narafičil , ty osle ! Román Gaius | je/oni/PPFP4--3-------- osel/osel/NNMS1-----A----- | ukazuje , že Henry Winterfeld psal už před půlstoletím lepší |
| Trutnovska z obce Mladé Buky , že v ulici Tmavá | je/být/VB-S---3P-AA---I osel/osel/NNMS1-----A----- | . „ Policisté díky místní znalosti vypátrali jeho majitele a |

# rožeň/rožnit/rozžehnout

| | | |
|---|---|---|
| NY . , O. LI . V . Y , | ROŽNI/rožeň/NNIP7-----A----- | , K. I . Š , Pa . Le . |
| to vím jistě . Z nářečí znám třeba slovo „ | rožni/rožeň/NNIP7-----A----- | “ . Ester Hozáková , 6 let , FrýdekMístek : |
| k paní Haleové , „ že s těmi hodinami , | rožni/rožeň/NNIP7-----A----- | a planetáriem máte doma malé muzeum . “ „ Jestli |
| jste naučila vy jeho ? Různé moravismy jako šufánek nebo | rožni/rožeň/NNIP7-----A----- | . Je nevěra tak zajímavá , že o ní teď |
| nás chodí dívat místní . Je tam obrovský krb s | rožni/rožeň/NNIP7-----A----- | , ve kterém se peče drůbež - od bažanta přes |
| je jen krátký úryvek z nabídky mas na grilu i | rožni/rožeň/NNIP7-----A----- | , které Farma nabízí . Mezi omáčkami , přílohami a |
| pro služebnictvo sestoupil do kuchyní , kde skomíral oheň pod | rožni/rožeň/NNIP7-----A----- | bez masa . Na zemi vedle stolu na porcování masa |
| pražsky , což rozhodně popírám ! Děti říkají fajne , | rožni/rožeň/NNIP7-----A----- | místo rozsviť , štrampliky místo punčocháče . Opakují slova , |
| , ale také teplá kuchyně s grily , rošty a | rožni/rožeň/NNIP7-----A----- | , kde gastronomické poklady před našima očima do zlatova dozrávaly |
| místnosti Dechem řádky staletí s krápníky sazí z ohňů pod | rožni/rožeň/NNIP7-----A----- | prasat či skopců hovoří centrální černá kuchyně , kam by |
| sklad plný všelikého harampádí a jen mohutné topeniště krbu s | rožni/rožeň/NNIP7-----A----- | a naproti stojící sporák byly důkazem , že stojím v |
| to pyšný . V Praze pořád používám slovo žufánek nebo | rožni/rožeň/NNIP7-----A----- | . A nerozumějí mi , “ prozradil . Když Okamura |
| , ale také teplá kuchyně s grily , rošty a | rožni/rožeň/NNIP7-----A----- | , kde gastronomické poklady před našima očima dozlatova dozrávaly . |

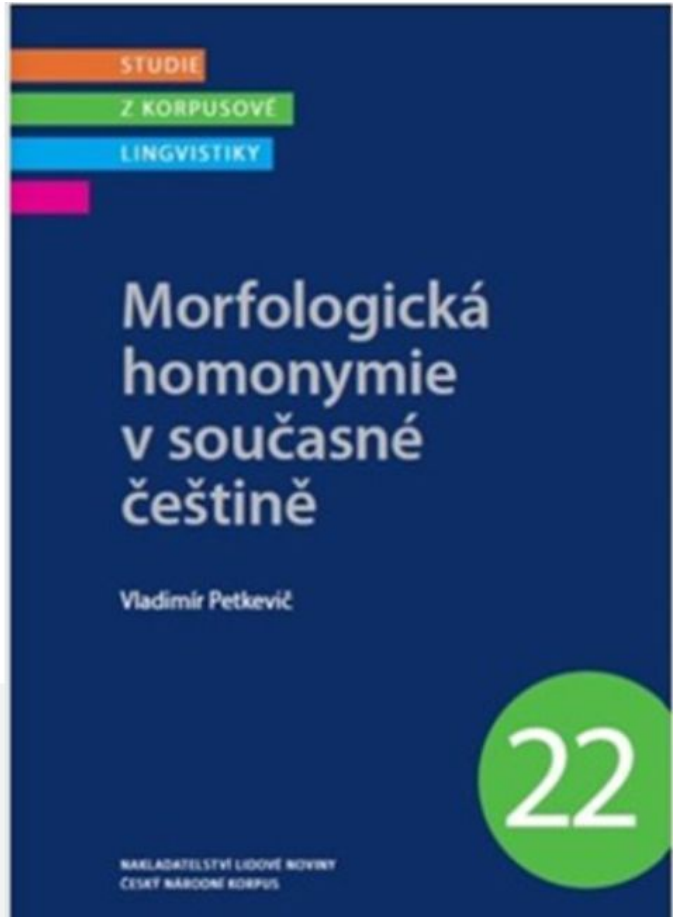
# Desambiguace

Statistická

Pravidlová (různé postupy - pravidla založená na syntaktických vztazích i na kolokacích)

Hybridní

# HOMONYMIE


STUDIE
Z KORPUSOVÉ
LINGVISTIKY
Morfologická homonymie v současné češtině
Vladimír Petkevič
22
NAKLADATELSTVÍ LIDOVÉ NOVINY
ČESKÝ NÁRODNÍ KORPUS

# Závěr

- Lemmatizace a tagging – užitečný nástroj
- Výsledky automatické analýzy je třeba prověřit s ohledem na záměr, který sleduje analýza dat
- Výsledky automatické analýzy lze zlepšovat
- Pokud pracujeme s výsledky automatické analýzy (s anotovaným korpusem), pak je třeba seznámit se i se způsoby automatické anotace a s řešením sporných / obtížných otázek.
- Pokud je desambiguace ve většině případů správně, pak s ní lze pracovat. Pokud je velmi špatná, je třeba hledat cesty, jak se bez ní obejít. Ty pak mohou být inspirací jak pro její zlepšení, tak pro uživatele neanotovaných korpusů.

# Vyzkoušejte

- Viděli jsme, že jedním z problémů, který ovlivňuje desambiguaci, je i tokenizace jednotek, které z lingvistického pohledu tvoří jedno textové slovo. Zamyslete se nad postupem, který by umožnil filtrovat případy, kdy slovní tvar *di* není substandardním imperativem 2. osoby sg. slovesa *jít*.
- Tvar *kolem* lze interpretovat jako substantivum, adverbium, nebo předložku. Pozorujte výsledky desambiguace tvaru *kolem* jako substantiva v korpusu SYN2020 v případě, že za tvarem *kolem* následuje tvar jména v genitivu a zamyslete se nad tím, jak by bylo možné chyby v desambiguaci filtrovat.
- V korpusu synv8 vyhledejte tvar *rožni* a pokuste se najít postup, jak filtrovat případy, kdy tvar je imperativem slovesa *rozsvítit*.

# Vyzkoušejte

Mezijazyková homonyma jsou jedním z problémů homonymie. Tak např. “***my***” je v češtině i angličtině slovo, v obou jazycích jde o zájmeno a v obou jazycích je frekventované. Přesto je žádoucí, aby vyskytne-li se v textu A slovo z jazyka B, bylo toto slovo správně desambiguováno (aby bylo správně rozpoznáno na všech úrovních automatické analýzy). Podívejte se do korpusu syn2020 na tvar “***my***”zobrazte jeho interpretace na úrovni morfologické značky a pokuste se zamyslet nad způsobem, jakým by bylo možné postupovat při snaze zlepšit desambiguaci. Porovnejte rozdíly ve značkování korpusů řady syn a korpusů webových dostupných přes Sketch Engine a pokuste se najít strategii pro řešení úkolu v korpusu cztenten.

(*Na to, jak blízké vztahy máme, tak přišli po dost dlouhé době, ale zatímco jsme my baby kecaly, chlapi navrtali sádroše na strop koupelny.*)

# Vyzkoušejte

Vyzkoušejte např. v BNC podobným způsobem najít případy, kdy tvar ***die*** není tvarem anglického slovesa.

Poměrně velmi zle vypadají výsledky desambiguace tvaru ***house*** v korpusu synv8. Navrhněte, jak byste postupovali, pokud byste chtěli najít skutečné doklady, kdy jde o neutrum pojmenovávající mládě husy.

# Tagsety v korpusech sketch engine: https://www.sketchengine.eu/tagsets/

- https://www.sketchengine.eu/tagset-reference-for-czech/

Some languages have more than one available POS tagset.

| LANGUAGES | | | |
|---|---|---|---|
| Amharic tagset | Arabic tagsets | Basque tagset | Bengali tagset |
| Bosnian tagset | Bulgarian tagsets | Burmese tagset | Catalan tagset |
| Chinese tagsets | Croatian tagset | Cundeelee Wangka tagset | Czech tagset |
| Danish tagsets | Dutch tagsets | English tagsets | Estonian tagsets |
| Finnish tagsets | French tagsets | German tagsets | Greek tagsets |

# Otázky

- Které tagsety používáte?
- Jaké jsou jejich výhody/nevýhody?
- Jaké jsou problémy automatické analýzy u jazyků typologicky odlišných od češtiny?

# Děkuji vám za pozornost