

MUNI

FI – Jazykové korpusy

CORE042

Pavel Rychlý

Kdo

- Pavel Rychlý
- Fakulta informatiky MU
- Centrum zpracování přirozeného jazyka
 - `nlp.fi.muni.cz`

Obsah

- jazykový korpus
- velikosti
- použití
- vytváření
- obohacování
- statistické zpracování
- publikace

Jazykový korpus

- kolekce textů přirozeného jazyka
- rozsáhlá
- v jednotném formátu

Obsah korpusu

- jazyk: např. angličtina, čeština, tisíce jiných
- poskytuje příklady autentického **užití jazyka**
- textový / mluvený
- podjazyky:
 - oborové (texty z webu FI)
 - žánrové (divadelní hry Shakepearra)
 - dobové (období husitství)

Velikosti korpusů

- čím větší, tím lepší
- často jsou limitem zdroje textů
 - Shakespeare už toho víc nenapíše
- první koprusy: 1 milion slov
 - příliš malé pro zajímavější výsledky
 - délka věty/slova, nejčastější slova
- nyní běžně miliardy slov
 - průměrná rychlost čtení je 125–225 slov za minutu
 - $200 * 60 * 18 = 216000$ slov za den (18 hodin)
 - 79 milionů za rok (365 dní)

Velikosti korpusů

- nyní běžně miliardy slov
 - průměrná rychlost čtení je 125–225 slov za minutu
 - $200 * 60 * 18 = 216000$ slov za den (18 hodin)
 - 79 milionů za rok (365 dní)
 - dost velká slovní zásoba
- pracujeme také s osmdesátimiliardovými korpusy
 - zhruba 1000 let čtení při 18 hodinách denně
- ChatGPT (2023)
 - trénováno na 300 miliardách slov (web, knihy, wikipedie, ...)

Použití



Použití: lingvistika, lexikografie

- místo introspekce sledujeme skutečná užití jazyka
- objektivní, reprezentativní
- porovnávání (sub)jazyků: délka věty/slova, nejčastější slova
- která slova zařadit do slovníku
- chování jednotlivých slov (kontexty, domény)

Použití: studium jazyka

- skutečná užití místo umělých příkladů
- identifikace důležitých slov/frází
- kontexty slov (kolokace, předložky)
- vytváření učebních materiálů

Použití: získávání informací

- získávání informací z textu
 - klíčová slova
 - fakta: pojmenované entity, kdo, kdy, kde
- získávání informací o autorech
 - stylové charakteristiky (délky slov/vět, ...)
 - identifikace autorství

Použití: trénování modelů

- čistý text je lepší
- velké jazykové modely
 - Slama – český/slovanský LLM typu GPT
- systémy strojového překladu

Spolupráce s jinými fakultami MU

- FF
 - Brněnský mluvený korpus
 - Kapradí, Kačenka, RapCor (francouzský rap)
 - Opravidlo 2.0 (evropský grant)
- PedF
 - Česko-německý paralelní korpus
- LF
 - latina v diagnózách
- FSS
 - Propaganda (průběžně přidávané texty)

Spolupráce s firmami

- Lexical Computing
 - strategický partner FI
 - corpus.tools
 - Sketch Engine (standard v lexikografii)
- Konica Minolta
- Seznam.cz

Vybrané případy použití

- Macmillan English Dictionary – výkladový slovník (2002)
- překladový slovník laoštiny, tagalštiny a urdštiny do angličtiny a korejštiny (2019)
- terminologická databáze Českého úřadu zeměměřického a katastrálního (2015)
- texty pro jazykový model strojového překladu do amharštiny (2017)
- databáze frází pro prediktivní psaní v mobilních zařízeních, různé jazyky (2014, 2021)
- kreativní vytváření názvů vznikajících značek, angličtina (2011)
- Estonský národní korpus – lingvistika, lexikografie (2013, 2017, 2019, 2021, 2023)

Vytváření



Vytváření: zdroje dat

- databáze dokumentů (doc, pdf, ...)
- datové sady (XML)
- novinové zprávy (RSS)
- web

Vytváření: stahování stránek z webu

- většinou největší zdroj
- snadno dostupný, pro libovoný jazyk
- crawler (SpiderLing)
 - prochází stránky, následuje odkazy
 - sleduje jazyk, výtěžnost (kolik textu ze stažených dat)
 - paralelní stahování z více serverů
 - slušné zacházení (nepřetěžuje)
- odstranění hlaviček, patiček, menu, reklamy, ...
- až několik miliard slov za týden

Vytváření: filtrování

- detekce jazyka (odstranění/rozdělení)
- detekce nežádoucího obsahu
- odstranění duplicit

Nežádoucí obsah

- typy: spam, generovaný obsah, šum, strojový překlad
- detekce
 - záleží na úhlu pohledu
 - copywriting nevadí pro studium jazyka, vadí pro získávání informací
- často je vidět až z výsledku
 - je potřeba identifikovat zdroj/důvod
 - opakovat zpracování

Obohacování



Obohacování: metadata

- záleží na zdroji dat
- autor (nebo adresa)
- datum publikace
 - umožňuje zachytit vývoj jazyka (neologismy, ...)
- automaticky:
 - jazyk
 - obor, žánr

Obohacování: tokenizace

- rozdělení textu do tokenů (pozic)
- token = základní jednotka korpusu
- většinou slovo, číslo, interpunkce
- někdy víceslovné: *New York, out of*
- někdy části slov: *don't* = do + n ' t

Obohacování: značkování

- morfologické
 - základní tvary
 - slovní druhy (podstatné jméno, sloveso, ...)
 - gramatické kategorie (rod, číslo, pád, ...)
- syntax
 - jmenné fráze
 - závislosti mezi slovy (modifikátor, předmět, ...)

Morfologické značkování

Universal Dependencies

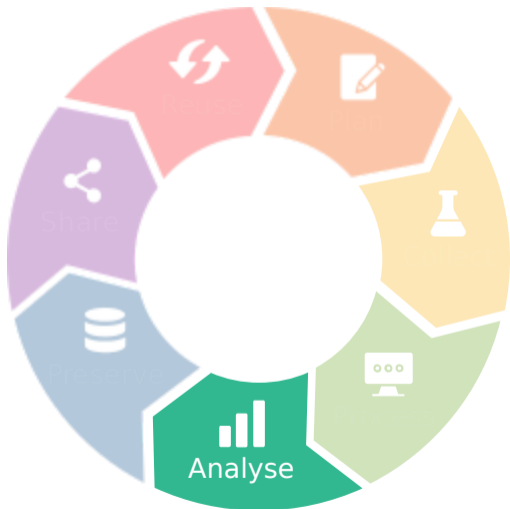
```
# newpar id = vesm9211-001-p7
# sent_id = vesm9211-001-p7s1
# text = Všechny tři světy si vzájemně trvale povídají a ovlivňují s
# orig_file_sentence vesm9211_001#8
Všechny    DET    Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|PronTy
tři        NUM    Case=Nom|Number=Plur|NumForm=Word|NumType=Card|NumVa
světy      NOUN   Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|Polarit
si         PRON   Case=Dat|PronType=Prs|Reflex=Yes|Variant=Short
vzájemně  ADV    Degree=Pos|Polarity=Pos
trvale     ADV    Degree=Pos|Polarity=Pos
povídají  VERB   Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Polarity=Po
a          CCONJ  _
ovlivňují VERB   Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Polarity=Po
se        PRON   Case=Acc|PronType=Prs|Reflex=Yes|Variant=Short
```

Morfologické značkování

Brněnské značky

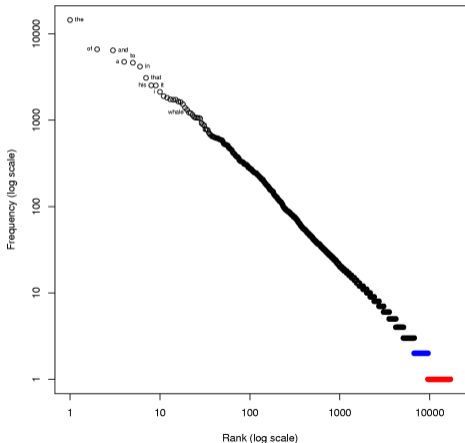
Z	z	k7c2
téměř	téměř	k6xMd1
tří	tři	k4xCgFnPc2
desítek	desítka	k4xNgFnPc2
smluv	smlouva	k1gFnPc2
upravujících	upravující	k2gFnPc2d1
vztahy	vztah	k1gInPc4
mezi	mezi	k7c7
oběma	dva	k4xCgInPc7
subjekty	subjekt	k1gInPc7
celního	celní	k2gMnSc2d1
soustátí	soustátí	k1gNnSc2
jsou	být	k5mItPp3nPaI
okamžitě	okamžitě	k6xMd1
vypověditelné	vypověditelný	k2gFnPc1d1
všechny	všechn	k3xUgFnPc1
.	.	kI

Statistické zpracování



Zipfův zákon

- $f * r = C$
- součin četnosti a pořadí v seznamu slov je konstantní



Statistické zpracování: kolokace

- význam slova určuje kontext
- kolokace = nejvýznamnější slova v okolí
- jaká slova následují?
 - záchodové XXX
 - tratoliště XXX
 - očitý XXX
 - polehčující XXX

Statistické zpracování: rámce slov

- gramatické relace
- definované pomocí morfologických značek

modifiers of "univerzita"	verbs with "univerzita" as subject	... of "univerzita"	verbs with "univerzita" as locale object
Karlův 109,910 ... Univerzity Karlovy	pořádat 540 ... univerzita pořádá	fakulta 98,065 ... fakulty Univerzity	studovat 14,413 ... studoval na univerzitě
Masarykův 65,947 ... Masarykovy univerzity	organizovat 129 ... univerzita organizuje	rektor 13,694 ... rektor Univerzity	vystudovat 6,812 ... vystudoval na univerzitě
západočeský 18,119 ... Západočeské univerzity v Plzni	spolupracovat 228 ... univerzita spolupracuje	houston 12,377 ... sam houston univerzity	přednášet 4,981 ... přednášel na univerzitě
jihocheský 19,113 ... Jihočeské univerzity	realizovat 94 ... univerzita realizuje	lešení 10,571 ... lešení univerzity rhode	vyučovat 2,700 ... vyučoval na univerzitě
ostravský 14,440 ... Ostravské univerzity	udělit 54 ... univerzita udělí	student 11,703 ... studentů Univerzity	učit 1,755 ... učí na univerzitě
slezský 12,014 ... Slezské univerzity v Opavě	zvat 237 ... univerzita zve	profesor 6,334 ... profesor univerzity	absolvovat 2,180 ... absolvoval na univerzitě
Mendelův 10,588 ... Mendelovy univerzity v Brně	otevřít 125 ... univerzita otevře	věda 4,819 ... Fakulty sociálních věd Univerzity Karlovy	působit 4,698 ... působil na univerzitě
zemědělský 13,368 ... České zemědělské univerzity	otevřít 131 ... univerzita otevřívá	studie 7,944 ... sociálních studií Masarykovy univerzity	promovat 490 ... promoval na univerzitě

Publikace



Publikace

- korpusy jsou velká data
- csTenTen17: 10 mld slov, 85 GB
 - zpracované pro vyhledávání 356 GB
- nejde posílat mailem
- nejde otevřít v editoru
- většina programů není schopna zpracovat

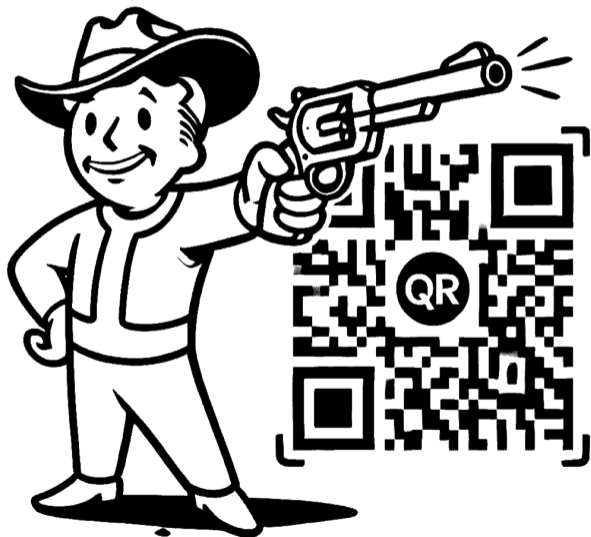
Publikace: datová úložiště

- LINDAT – CLARIAH-CZ
 - CLARIN, DARIAH
 - zajišťuje dostupnost, FAIR
 - csTenTen17: <http://hdl.handle.net/11234/1-4835>
- Licence
 - většinou nemáme autorská práva
 - často nejsme schopni je získat
 - NLP Centre Web Corpus License Agreement

Publikace: webová rozhraní

- místo stahování dat využíváme službu
- web je univerzální – dostupné odkudkoliv
- rychlé vyhledávání
- nejsme toklik omezení autorskými právy
 - nejsou přístupné celé texty
- Sketch Engine: <https://app.sketchengine.eu>
 - přístup přes univerzitní učo a heslo (*Institutional login*)

KvIS!



Souhrn

- jazykové korpusy poskytují reprezentativní užití jazyka
- často vytvářené z webu
- velká data
- studenti/zaměstnanci MU mají přístup ke Sketch Engine

Chcete více?

- obor **Digitální lingvistika** (Joint Master Degree)
 - 2 roky, 1 semestr v zahraničí
- IB047 Úvod do korpusové lingvistiky a počítačové lexikografie
 - bakalářský předmět na FI (jaro)
- PA153 Natural Language Processing
 - magisterský předmět na FI (v angličtině, podzim)
- **PV173** Seminář zpracování přirozeného jazyka
 - práce v laboratoři NLP
 - **kredity každý semestr**