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 přírozeného 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 Shakepeara)
  - 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 stovky milionů 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 stovky milionů 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
- dostupné jsou i giga-korpusy
  - více než miliarda slov
  - zhruba 50 let čtení při 4 hodinách denně
- ChatGPT
  - trénováno na 300 miliardách slov (web, knihy, wikipedie, …)

Použití

Použití: ligvistika, 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

Spolupráce s jinými fakultami MU

- FF
  - Brněnský mluvený korpus
  - Kapradí, Kačenka
  - RapCor (francouzský rap)
- 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

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, reklany, …
- 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

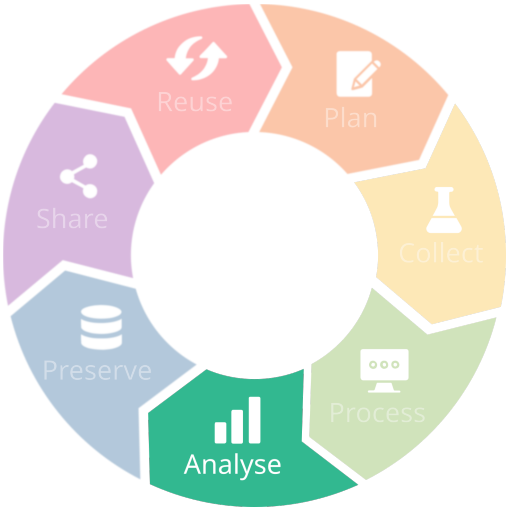
```
# newpar id = vesm9211-001-p7
# sent_id = vesm9211-001-p7s1
# text = Všechny tři světy si vzájemně trvale povídají a ovlivňují s
# orig_file_sentence vesm9211_001#8
Všechny    DET   Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|PronTy
tři        NUM   Case=Nom|Number=Plur|NumForm=Word|NumType=Card|NumVa
světy      NOUN  Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur|Polari
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

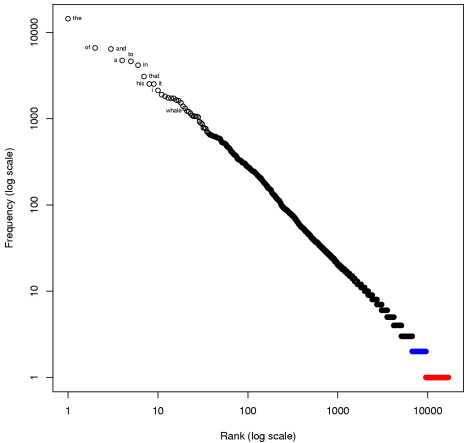
Z	z	k7c2
téměř	téměř	k6xMd1
tří	tři	k4xCgFnPc2
desítek	desítka	k4xNgFnPc2
smluv	smlouva	k1gFnPc2
upravujících	upravující	k2gFnPc2d1
vztahy	vztah	k1gInPc4
mezi	mezi	k7c7
oběma	dva	k4xCgInPc7
subjekty	subjekt	k1gInPc7
celního	celní	k2gMnSc2d1
soustátí	soustátí	k1gNnSc2
jsou	být	k5mItPp3nPaI
okamžitě	okamžitě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

Statistické zpracování



Zipfův zákon

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

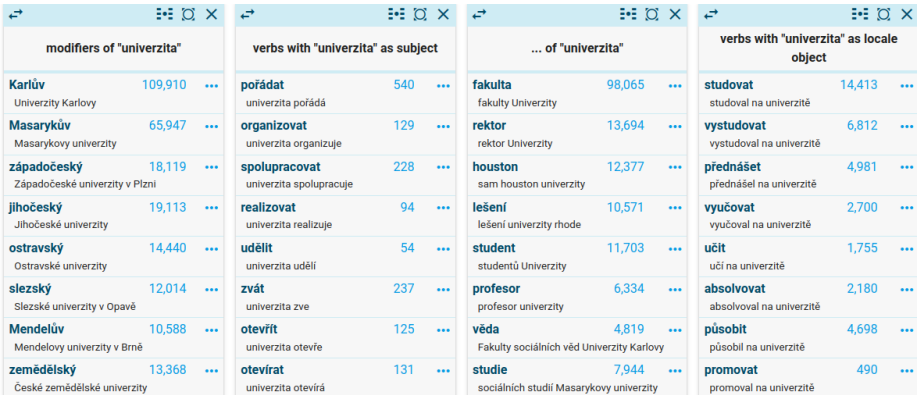


Statistické zpracování: kolokace

- 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"		
Karlův	109,910	•••
Univerzity Karlovy		
Masarykův	65,947	•••
Masarykovy univerzity		
západočeský	18,119	•••
Západočeské univerzity v Plzni		
jihočeský	19,113	•••
Jihočeské univerzity		
ostravský	14,440	•••
Ostravské univerzity		
slezský	12,014	•••
Slezské univerzity v Opavě		
Mendelův	10,588	•••
Mendelovy univerzity v Brně		
zemědělský	13,368	•••
České zemědělské univerzity		

verbs with "univerzita" as subject		
pořádat	540	•••
univerzita pořádá		
organizovat	129	•••
univerzita organizuje		
spolupracovat	228	•••
univerzita spolupracuje		
realizovat	94	•••
univerzita realizuje		
udělit	54	•••
univerzita udělí		
zvát	237	•••
univerzita zve		
otevřít	125	•••
univerzita otevře		
otevírat	131	•••
univerzita otevírá		

... of "univerzita"		
fakulta	98,065	•••
fakulty Univerzity		
rektor	13,694	•••
rektor Univerzity		
houston	12,377	•••
sam houston univerzity		
lešení	10,571	•••
lešení univerzity rhode		
student	11,703	•••
studentů Univerzity		
profesor	6,334	•••
profesor univerzity		
věda	4,819	•••
Fakulty sociálních věd Univerzity Karlovy		
studie	7,944	•••
sociálních studií Masarykovy univerzity		

verbs with "univerzita" as locale object		
studovat	14,413	•••
studoval na univerzitě		
vystudovat	6,812	•••
vystudoval na univerzitě		
přednášet	4,981	•••
přednášel na univerzitě		
vyučovat	2,700	•••
vyučoval na univerzitě		
učit	1,755	•••
učí na univerzitě		
absolvovat	2,180	•••
absolvoval na univerzitě		
působit	4,698	•••
působil na univerzitě		
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*)

Souhrn

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