Digital humanities

Analýza textů

Jindřich Marek

Kontext analýzy textu

- zpracování přirozeného jazyka (NLP)
- strojové učení (ML)
- statistická analýza

Úvod do analýzy textu

- získávání textu
- vytěžování textu
- nástroje
- příklady

Získávání textu

- webscraping
- OCR v digitálních knihovnách
- rozpoznávání řeči
- HTR: OCR pro rukopisné texty
- …

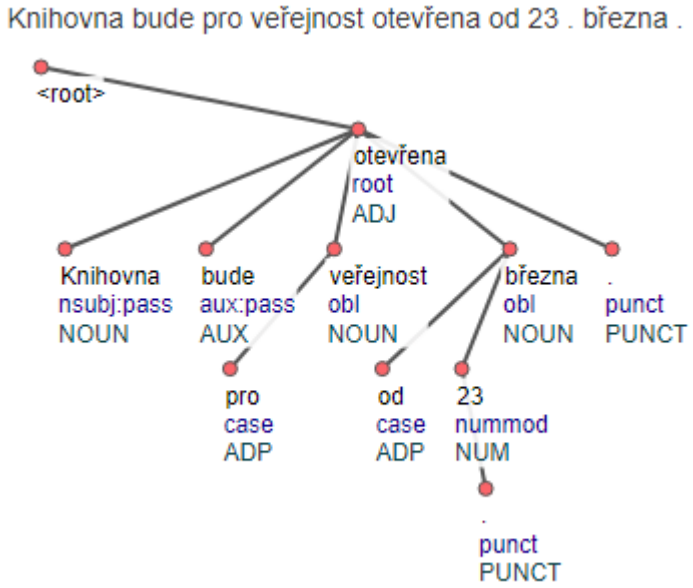
Vytěžování textu

- extrakce vzorů, trendů a vztahů
  - z rozsáhlých textových souborů dat

- tokenizace
  - rozdělení textu na slovní tvary a interpunkční znaménka
- lemmatizace
  - určení základních tvarů slov

Lemmatizace

- příklad: UDPIpe
  - https://lindat.mff.cuni.cz/services/udpipe/



text = Knihovna bude pro veřejnost otevřena od 23. března.

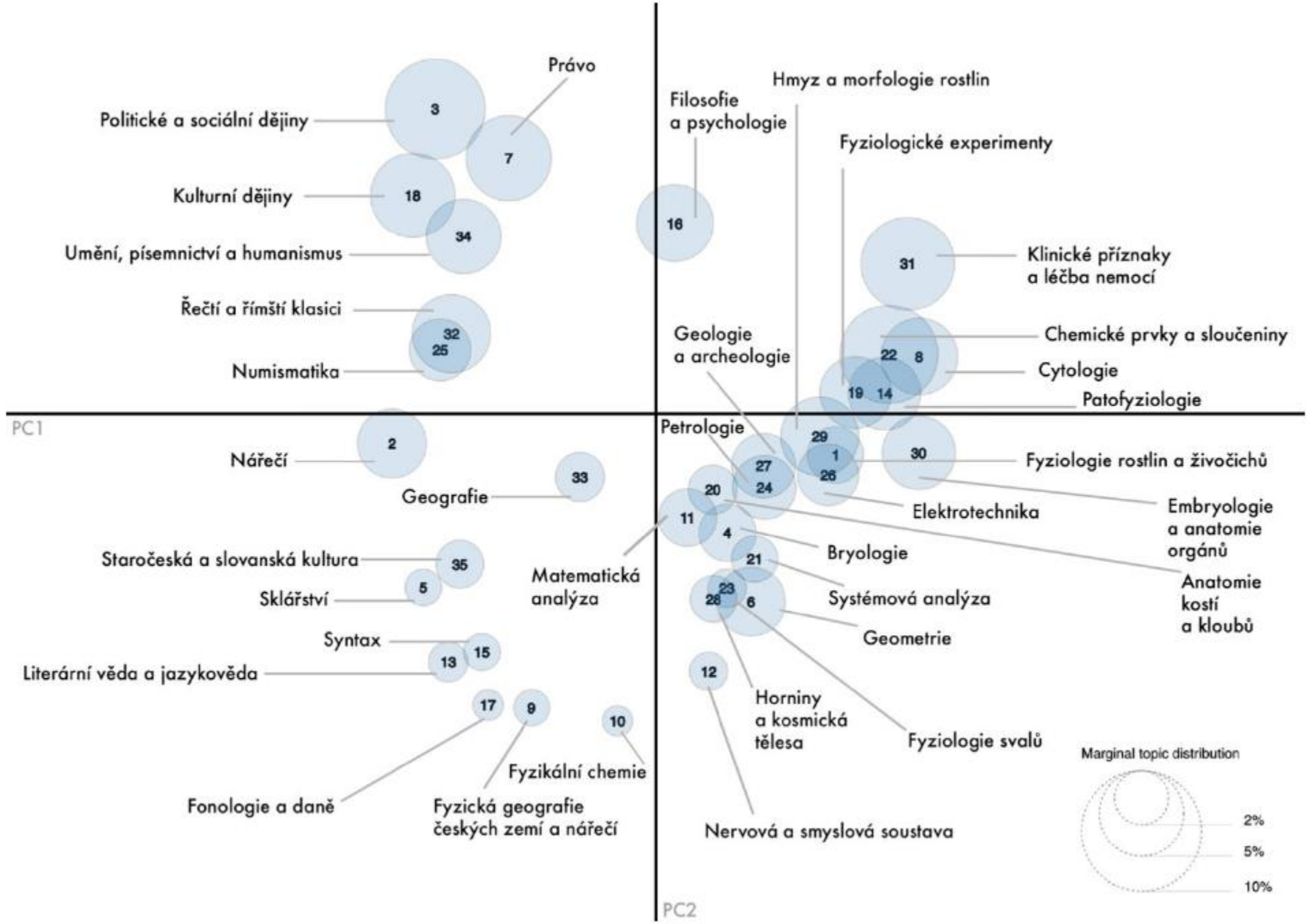
1	Knihovna	knihovna	NOUN	NNFS1-----A----	Case=Nom|Gender=Fem|Number=Sing|Polarity=Pos	5	nsubj:pass	_	TokenRange=271:279
2	bude	být	AUX	VB-S---3F-AAI--	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Fut|VerbForm=Fin|Voice=Act	5	aux:pass	_	TokenRange=280:284
3	pro	pro	ADP	RR--4----------	AdpType=Prep|Case=Acc	4	case	_	TokenRange=285:288
4	veřejnost	veřejnost	NOUN	NNFS4-----A----	Case=Acc|Gender=Fem|Number=Sing|Polarity=Pos	5	obl	_	TokenRange=289:298
5	otevřena	otevřený	ADJ	VsQW----X-APP--	Aspect=Perf|Gender=Fem,Neut|Number=Plur,Sing|Polarity=Pos|Variant=Short|VerbForm=Part|Voice=Pass	0	root	_	TokenRange=299:307
6	od	od	ADP	RR--2----------	AdpType=Prep|Case=Gen	9	case	_	TokenRange=308:310
7	23	23	NUM	C----------	NumForm=Digit|NumType=Card	9	nummod	_	SpaceAfter=No|TokenRange=311:313
8	.	.	PUNCT	Z:----------	_	7	punct	_	TokenRange=313:314
9	března	březen	NOUN	NNIS2-----A----	Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing|Polarity=Pos	5	obl	_	SpaceAfter=No|TokenRange=315:321
10	.	.	PUNCT	Z:----------	_	5	punct	_	TokenRange=321:322

Analýza textu

- modelování témat
- analýza sentimentu
- analýza textových sítí, rozpoznávání pojmenovaných entit
- stylometrie
- digitální edice textů
- vizualizace textu
- korpusová lingvistika

Modelování témat

- cílem je odhalit skryté tematické struktury v souboru dokumentů
- latentní Dirichletova alokace (LDA)
- příklad: publikační činnost České akademie věd a umění 1890-1910 (Filip Kersch) >>>


Právo
Politické a sociální dějiny
3
7
Kulturní dějiny
18
34
Umění, písemnictví a humanismus
Řečtí a římští klasici
32
25
Numismatika
Filosofie
a psychologie
Hmyz a morfologie rostlin
Fyziologické experimenty
16
31
Klinické příznaky
a léčba nemocí
Geologie
a archeologie
Chemické prvky a sloučeniny
22
8
Cytologie
19
14
Patofyziologie
PC1
Petrologie
2
Nářečí
29
1
30
Fyziologie rostlin a živočichů
33
Geografie
27
26
20
24
Elektrotechnika
Embryologie
a anatomie
orgánů
11
4
Bryologie
Staročeská a slovanská kultura
35
Matematická
analýza
21
Anatomie
kostí
a kloubů
5
Sklářství
23
28
6
Systémová analýza
Geometrie
Syntax
15
13
Literární věda a jazykověda
12
Horniny
a kosmická
tělesa
17
9
10
Fyzikální chemie
Fyziologie svalů
Fonologie a daně
Fyzická geografie
českých zemí a nářečí
Nervová a smyslová soustava
Marginal topic distribution
2%
5%
10%
PC2

Analýza sentimentu

- určení sentimentu nebo emocionálního tónu vyjádřeného v textu
- základ: pozitivní, negativní nebo neutrální

Analýza textových sítí

- mapování vztahů mezi entitami nebo pojmy reprezentovanými v textových datech

- speciální případ: rozpoznávání pojmenovaných entit (NER)
  - https://lindat.mff.cuni.cz/services/nametag/

Želivský klášter má nově uspořádanou knihovnu, historický mobiliář byl zničený

Aktualizace: 22.03.2024 11:02 Vydáno: 22.03.2024, 11:02

Opat Tadeáš Róbert Spišák ve zrekonstrované knihovně želivského premonstrátského kláštera, 19. března 2024 Želiv, Pelhřimovsko. Knihovna bude pro veřejnost otevřena od 23. března. CTK/Pavlíček Luboš

Želiv (Pelhřimovsko) - Želivský klášter má nově uspořádanou historickou knihovnu, v níž je okolo 25.000 svazků napsaných do roku 1850. V sále barokního konventu se zachovaly fresky, ale původní vybavení knihovny bylo zničené. Premonstráti tam nechali osadit nový mobiliář. Od 13. dubna budou přístupné další nové klášterní expozice v opravené budově Staré prelatury, přiblíží historii kláštera a také někdejší bydlení opata. ČTK to řekli zástupci kláštera.

A Raw Output | Highlighted Output

Želivský klášter má nově uspořádanou knihovnu, historický mobiliář byl zničený
Aktualizace: 22.03.2024 11:02 Vydáno: 22.03.2024, 11:02
foto
Opat Tadeáš Róbert Spišák ve zrekonstrované knihovně želivského premonstrátského kláštera, 19. března 2024 Želiv, Pelhřimovsko. Knihovna bude pro veřejnost otevřena od 23. března. ČTK/Pavlíček Luboš
Želiv (Pelhřimovsko) - Želivský klášter má nově uspořádanou historickou knihovnu, v níž je okolo 25.000 svazků napsaných do roku 1850. V sále barokního konventu se zachovaly fresky, ale původní vybavení knihovny bylo zničené. Premonstráti tam nechali osadit nový mobiliář. Od 13. dubna budou přístupné další nové klášterní expozice v opravené budově Staré prelatury, přiblíží historii kláštera a také někdejší bydlení opata. ČTK to řekli zástupci kláštera.

Fotogalerie

Opat Tadeáš Robert Spišák zavzpomínal na to, jak po příchodu do Želiva před 20 lety prováděl návštěvníky klášterem. V refektáři, což je jídelna, zdůrazňoval, že nad ním je knihovna, duchovní pokrm. Zároveň tehdy říkal, že knihovna je zavřená. "Protože nábytek byl za komunismu zničen, spálen a ty knihy jsou jenom tak na hromadách a potřebujeme je katalogizovat," řekl ČTK.

Knihovna teď má nový mobiliář tvořený volně stojícími betonovými podpěrami a ocelovými policemi, do nichž jsou cenné knihy převážně v kožených vazbách pečlivě srovnané. "Rozhodli jsme se po poradě s památkovým úřadem, že nepůjdeme do nějakých replik, protože se nám nic z té historické staré knihovny nezachovalo," uvedl opat.

Mobiliář navrhla kancelář Šépka architekti. Při odlévání betonových dílů byl použitý štěrk a kamínky z řeky Želivky. S uspořádáním knih a jejich katalogizací pomáhají studenti Univerzity Palackého v Olomouci pod vedením Jany Oppeltové, která vyučuje na katedře historie Filozofické fakulty.

"Klášter nemůže fungovat bez knih," řekla Oppeltová. Uvedla, že želivská knihovna měla být odrazem světa, proto tam kromě teologické literatury jsou knihy z různých oborů. Psané jsou hlavně německy, latinsky a česky. Jsou tam i středověké a barokní rukopisy a prvotisky vytištěné mezi roky 1450 až 1500. Po zrušení kláštera komunisty byly knihy odvezené na Strahov. Z větší části se dochovaly, ale byly pomíchané s knihami z dalších klášterů, uvedla.

Knihovna bude k dispozici odborné veřejnosti a při zvláštních příležitostech také návštěvníkům. Opravená Stará prelatura je součástí prohlídkového okruhu kláštera. Expozice v ní přibližují tamní archeologické nálezy, vývoj kláštera od jeho založení v roce 1139 i to, jak mohlo vypadat bydlení opata přibližně na začátku 20. století a jaká byla jeho role v klášteře a ve společnosti. Vystavené tam jsou také památky doposud uložené v depozitářích.

V budově blízko kostela byly naposled byty, pak byla kvůli špatnému stavu delší čas prázdná. Opravy skončily vloni. Klášter teď má rovněž nové sociální zázemí pro návštěvníky. Na investice skoro za 50 milionů korun dostala kanonie přes 40 milionů korun z programu IROP.

Klášter zrušili komunisté v roce 1950. Šest let tam byl internační tábor pro kněze a řeholníky, následně psychiatrická a protialkoholní léčebna. Premonstráti se mohli vrátit počátkem 90. let, budovy postupně opravují. Areál, který je národní kulturní památkou, si loni prohlédlo 17.000 návštěvníků.

Stylometrie

- analýza stylu nebo charakteristik psaní autorů na základě jazykových znaků, jako je slovní zásoba, syntax a interpunkce

- atribuce autorství atd.

- příklad: QuitaUp
  - https://korpus.cz/quitaup/

QuitaUp

Choose a file

Browse zeliv.txt

Upload complete

Language

Czech

Units

Word forms (case insensitive)

Ignore punctuation

Preview Results About

Results of individual texts

Index	Value
Tokens	480
Types	338
TTR	0.704
h-point (h)	6
Hapaxes	272
Hapax legomena percentage	0.567
Entropy (H)	8.081
Verb distance (VD)	9.795
Activity (Q)	0.388
Descriptivity (D)	0.612
Average Token length (ATL)	5.631
Thematic concentration (TC)	0.0777
Secondary thematic concentration (STC)	0.1288
Moving Average Type-Token Ratio (MATTR, L=100)	0.87
Moving Average Type-Token Ratio (MATTR, L=500)	
zTTR: Normalized TTR	0.855
Moving Average Morphological Richnes (MAMR, L=100)	0.0744
Moving Average Morphological Richnes (MAMR, L=500)	

Results (.csv)

Thematic words and their weights

Word	POS	TW (primary TC)	TW (secondary TC)
kláštera	NOUN	0.0389	0.0498
knihovna	NOUN	0.0389	0.0498
klášter	NOUN	NA	0.0146
tam	ADV	NA	0.0146

For POS tags see UD documentation

Vertical format

Digitální edice textů

- vědecké edice literárních nebo historických textů
  - anotace, textové varianty a další metadata pro vědeckou analýzu a interpretaci

- příklad: Vokabulář webový (edice)
  - https://vokabular.ujc.cas.cz/moduly/edicni/

[O ženě zlobivé]

Knihovna Národního muzea v Praze (Praha, Česko), sign. II F 8, 187rv. Editor Pečírková, Jaroslava. Ediční poznámka

187r–187v

Číslo stránky: Přejít

[187r]

Pakli kto má ženu zlobivú,

ale kup jí sukni novú,

a u vetchéj kaž vše dni choditi,

budeť sě daleko měně zlobiti.

Pakliť bude láti druhé,

a ty jí kup třěvíce nové;

kažiž vetché[a] dievce dáti,

vždyť již bude méně láti.

Pakli k ní který hněv máš,

kup jí měšec, nuož i nový pás,

a čímť bude viece rotiti,

máš ji vždy dary krotiti,

a čímť bude viece láti,

a ty jí máš vždy dar dáti.

Pakliť jest nemilá proč,

Vizualizace textů

- vizuální znázornění textových dat
  - zkoumání a interpretace vzorců a vztahů
  - mračna slov (wordcloud), síťové grafy a heatmapy

- příklady (po lemmatizaci):
  - vánoční poselství prezidenta Miloše Zemana 2022 (poslední)
  - novoroční projev prezidenta Petra Pavla 2024 (první)

republika
rok
zvolený
zájem
předseda
státní
inflace
český
vláda
prezident
chtít
země
vztah
rozpočet
ekonomický
první
druhý
deficit
strana
cena
volba
agrese
řada
politický
demokracie
doba
ukrajina
pokus
sazba
dva
růst
milý
tři
faktor
banka
snažit
sociální
samospráva
procento
aktuální
velký
ruský
národní
politika
naposledy
zejména
zvyšování
celý
událost
reforma
krize
třetí
pět
stát
jaderný
náklad
úrokový
bankovní
podpora
podmínka
starobní
armáda
deset
koruna
úcta
poselství
elektrárna
revoluce
důsledek
suverenita
vstoupit
dostat
politik
krajský
setkávat
práce
problém
významný
příklad
přispět
úvěr
zahraniční
poprvé
devadesát
zvyšovat
krásný
důchod
export
energetický
malý
nula
jeden
osmdesát
izraelský
množství
podporovat
přijmout
energie
bezpečnost
miliarda
zdravý
rovněž
korektní
třicet
tlak
právě
život
den
vánoční
nyní
izrael
rámec
část
deal
člen
čest
prožít
odborník
koneckonců
agentura
nikoliv
alespoň
oprávnění
guvernér
vchod
ekonom
uvést
povolenka
voják
praha
upozornit
oponent
přejít
ukazovat
splnit
sousední
oblast
německo
tvrdý
zvýšit
subjekt
zdražování
mladý
evropa
výbor
sektor
nejenom
významně
bojující
novela
spalovací
bezpečnostní
rozšířit
příjmový
domácnost
ztráta
vládní
snižovat
kniha
bankrot
prominout
centrální
tehdy
výrazně
hlavní
dovolit
promítat
objekt
projít
celosvětový
fanatizmus
kritický
zabývat
základ
švejnar
ohrožený
vznik
využít
surovina
ruka
devět
šílený
spoluobčan
miloš
emisní
moci
rusko
vážený
číst
slib
řídit
opomíjený
jaro
říci
září
totiž
volit
rozdělit
bezesporu
občas
řešit
sto
shrnout
profesor
přimět
ohlas
pomoc
svět
dlouhý
jít
relativně
plně
věřit
složit
účet
rada
úsilí
řád
týkat
mluvit
dnešní
území
čas
osm
obvykle
prosit
levný
pozdě
situace
zeman
zelený
brzy
green
motor
dát
omezit
zakázat
trápit
auto
muž
česky
objem
stoupenec
podesáté
uplynulý
příliš
okupace
hájit
úrok
peníze
článek
ustoupit
návrh
služba
opustit
podesátý
vetovat
unie
uhelný
svobodný
závěr
energium
sedm
zákonný
povinnost
uspět
cesta
dodávka
názor
činit
věnovat
pokládat
slovo
federace
silný
daňový
stavba
lány
zákon

rok
moci
země
svět
politika
život
mladý
velký
stát
prezident
vláda
rodina
problém
přesto
evropský
člověk
přát
řešení
budoucnost
změna
příležitost
občan
možnost
společnost
očekávání
obrátit
myslet
veřejný
petr
pavel
strach
chvíle
unie
společný
ekonomický
potíž
pocit
zdraví
strana
projev
dařit
okolí
den
nechat
opřít
dělat
ruka
jednoduchý
připravený
konkrétní
tragický
důvěra
ekonomika
událost
dostat
hlavně
český
zdát
najít
dnešní
kontrola
naděje
diskuse
přinášet
vstup
politik
novoroční
právě
opatření
potřebovat
silný
domácí
vážený
generace
plán
viník
vědět
naplnění
růst
úkol
věc
vize
brzy
letos
závazek
štěstí
připomínat
rád
připomenout
zároveň
inflace
ohlédnout
populární
uzavřený
životní
prostředí
řešit
doba
spoluobčan
měna
politický
situace
současný
snaha
bezpečí
snažit
chtít
opozice
radost
očekávat
udělat
poprvé
záruka
začít
možný
mnohý
deset
pomáhat
tragédie
rozepře
jít
žít
čas
vůle
plný
zlo
zpráva
seznam
krok
tolik
pět
vnímat
pravomoc

Korpusová lingvistika

- systematická analýza rozsáhlých souborů textů (korpusů)
  - studium jazykových vzorců, užívání a variability jazyka

- příklad: Český národní korpus
  - https://korpus.cz/

Naše aplikace

Manuály k našim nástrojům s jejich stručným popisem jsou k dispozici na jednotlivých stránkách:

KonText	webové rozhraní pro práci s korpusy ČNK (přejít na aplikaci; API)
Slovo v kostce	agregátor slovních profilů (přejít na aplikaci)
SyD	nástroj pro korpusový průzkum variant (přejít na aplikaci)
Morfio	nástroj pro analýzu slovotvorných vztahů (přejít na aplikaci)
KWords	aplikace pro extrakci klíčových slov (přejít na aplikaci)
Treq	databáze překladových ekvivalentů (přejít na aplikaci; API)
Pro školy	stránka s korpusovými cvičeními pro výuku jazyka na ZŠ a SŠ (přejít na stránku)
Calc	korpusová kalkulačka umožňující snadno statisticky vyhodnotit výsledky hledání (přejít na aplikaci)
Lists	prohlížeč frekvenčních seznamů slov z hlavních korpusů ČNK (přejít na aplikaci)
Mapka	mapová aplikace pro korpusy mluvené češtiny (přejít na aplikaci)
KorpusDB	databáze slovních tvarů a lemmat doložených v korpusech ČNK (přejít na aplikaci)
QuitaUp	nástroj pro výpočet stylometrických ukazatelů z textu (přejít na aplikaci)
GramatiKat	nástroj pro výzkum gramatických kategorií (přejít na aplikaci)
Akalex	aplikace pro výzkum slovní zásoby akademické češtiny (přejít na aplikaci)
Alpha	překladač dotazů z přirozeného jazyka do CQL (přejít na aplikaci)

Případné potíže a nejasnosti v používání nástrojů ČNK je také možné kdykoli konzultovat v online foru Poradna ČNK (registrace nutná).

Word at a Glance

Čeština

stále

pořád

furt

Search

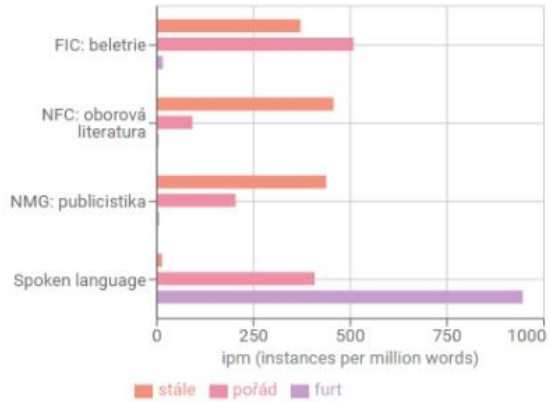
Frequency information

Basic characteristics

		occur. per mil. words	freq. band
1.	entered form: stále lemma: **stále** part of speech: [adverb]	438.3	★★★★☆
2.	entered form: pořád lemma: **pořád** part of speech: [adverb]	147.84	★★★★☆
3.	entered form: furt lemma: **furt** part of speech: [adverb]	3.59	★★☆☆☆

Source: syn_v8

Frequency according to a text type



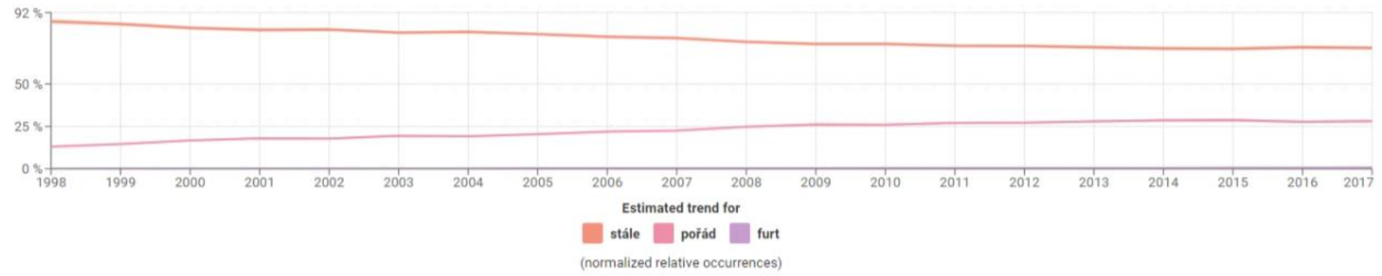
Source: syn2015 + oral_v1, more detailed information:
typy textů v SYN2015 (KonText), výsledky v ORAL verze 1 (KonText)

Written language ?

Collocations

Source: syn2015, more detailed information: Kolokace v SYN2015 (KonText), Kolokace v SYN2015 (KonText), Kolokace v SYN2015 (KonText)

Compared frequencies of the words over time



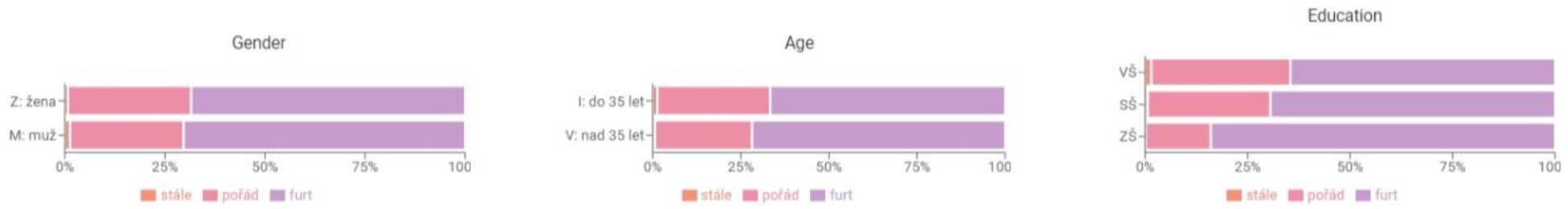
Source: syn_v7, more detailed information: KonText, KonText

Similarly used words

Source: syn_v8

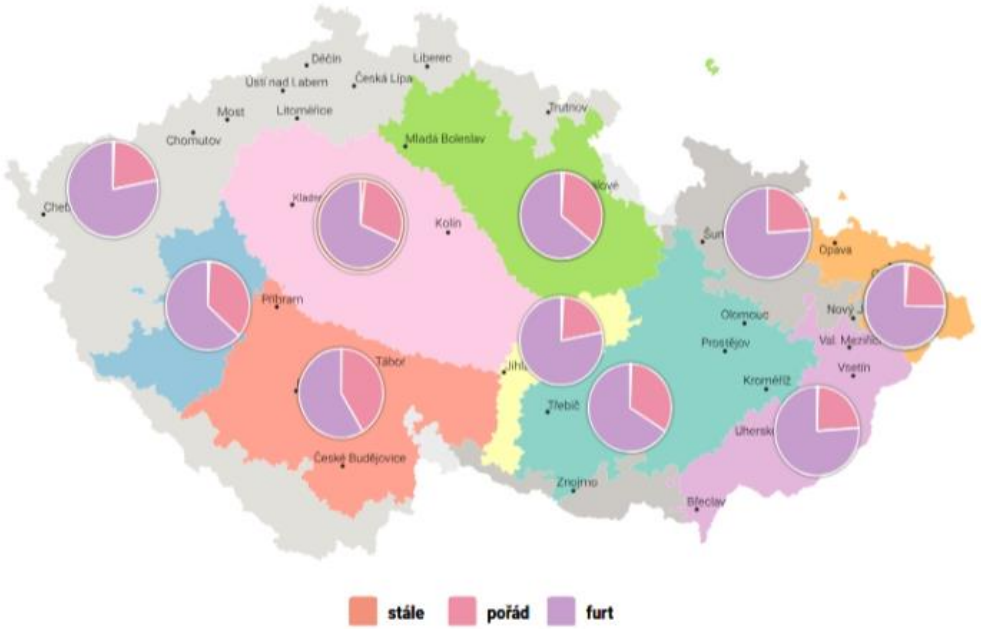
Spoken language

Frequency comparison



Source: oral_v1, more detailed information: KonText, KonText, KonText, KonText, KonText, KonText, KonText, KonText, KonText

Areas according to the traditional dialect division



Source: oral_v1, more detailed information: KonText, KonText, KonText

Nástroje (příklady)

- Voyant Tools
  - https://voyant-tools.org/
- R: tokenizers
  - https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html

Analýza anglického textu v R

- zprávy o stavu Unie amerických prezidentů
  - https://programminghistorian.org/en/lessons/basic-text-processing-in-r

Charakteristická slova: Zpráva o stavu Unie

William J. Clinton; 1993; deficit; propose; incomes; invest; decade
William J. Clinton; 1994; deficit; renew; ought; brady; cannot
William J. Clinton; 1995; ought; covenant; deficit; bureaucracy; voted
William J. Clinton; 1996; bipartisan; gangs; medicare; deficit; harder
William J. Clinton; 1997; bipartisan; cannot; balanced; nato; immigrants
William J. Clinton; 1998; bipartisan; deficit; propose; bosnia; millennium
William J. Clinton; 1999; medicare; propose; surplus; balanced; bipartisan
William J. Clinton; 2000; propose; laughter; medicare; bipartisan; prosperity
George W. Bush; 2001; medicare; courage; surplus; josefina; laughter
George W. Bush; 2002; terrorist; terrorists; allies; camps; homeland
George W. Bush; 2003; hussein; saddam; inspectors; qaida; terrorists
George W. Bush; 2004; terrorists; propose; medicare; seniors; killers
George W. Bush; 2005; terrorists; iraqis; reforms; decades; generations
George W. Bush; 2006; hopeful; offensive; retreat; terrorists; terrorist
George W. Bush; 2007; terrorists; qaida; extremists; struggle; baghdad
George W. Bush; 2008; terrorists; empower; qaida; extremists; deny
Barack Obama; 2009; deficit; afford; cannot; lending; invest
Barack Obama; 2010; deficit; laughter; afford; decade; decades
Barack Obama; 2011; deficit; republicans; democrats; laughter; afghan
Barack Obama; 2012; afford; deficit; tuition; cannot; doubling
Barack Obama; 2013; deficit; deserve; stronger; bipartisan; medicare
Barack Obama; 2014; cory; laughter; decades; diplomacy; invest
Barack Obama; 2015; laughter; childcare; democrats; rebekah; republicans
Barack Obama; 2016; laughter; voices; allies; harder; qaida

Donald J. Trump; 2017; allies; billions; borders; incredible; jenna
Donald J. Trump; 2018; heroes; isis; terrorists; deserve; kenton
Donald J. Trump; 2019; decades; brave; confront; defend; herman
Donald J. Trump; 2020; sanctuary; unemployment; aliens; ellie; arrested
Joseph R. Biden; 2021; applause; pandemic; consensus; republicans; vaccinated
Joseph R. Biden; 2022; pandemic; putin; ukrainian; allies; announcing
Joseph R. Biden; 2023; medicare; bipartisan; decades; inflation; seniors
Joseph R. Biden; 2024; predecessor; wealthy; bipartisan; corporations; defend