

Digital humanities

Analýza textů

Jindřich Marek

Kontext analýzy textu

- co je to text
- strojové učení (ML)
- zpracování přirozeného jazyka (NLP)
- statistická analýza

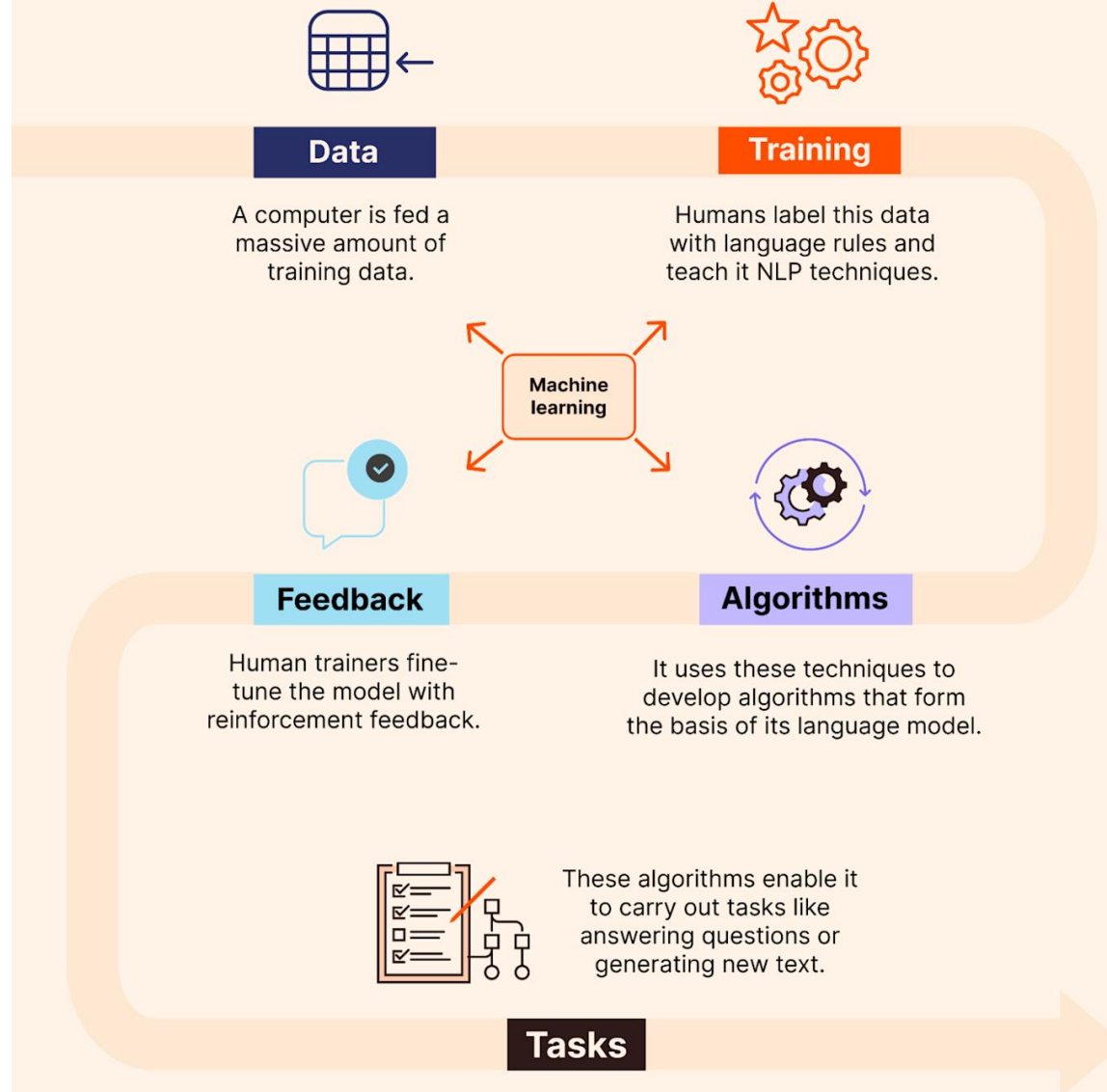
Úvod do analýzy textu

- získávání textu
- vytěžování textu
- nástroje

Získávání textu

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

How does natural language processing work?



Natural language processing (NLP) tasks

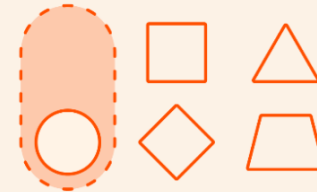
NLP is the process through which AI is taught to **understand the rules and syntax of language**, programmed to **develop complex algorithms to represent those rules**, and then made to **use those algorithms to carry out specific tasks** like these.



Language generation



Answering questions



Text classification



Sentiment analysis



Machine translation

Vytěžování a analýza textu

- extrakce vzorců, trendů a vztahů
 - z rozsáhlých textových souborů dat
- korpusová lingvistika

NLP: příprava

- normalizace vstupu
 - převod na Unicode, odstranění neviditelných znaků, příp. konverze velkých/malých písmen
- rozdělení textu do vět (sentence segmentation)
- tokenizace: rozdělení vět na tokeny (slova, čísla, interpunkci)

NLP: jádro

- morfologická analýza
 - lemmatizace: určení základních (lexikálních) tvarů slov
 - nebo stemming: určení kořenů slov, oddělení předpon a přípon; nižší kvalita
 - určení slovních druhů a tvarů slov (POS-tagging)
- syntaktická analýza
 - zjištění syntaktických vztahů (dependency parsing)
 - alternativně určení frázové struktury (constituency parsing)
- filtrace stop-slov (volitelné, podle úkolu)
 - vypuštění málo významových slov, která by znepřehledňovala analýzu
 - příklady českých stop-slov: aby, ale, což, ...
 - po tokenizaci a často po lemmatizaci (nebo před ní)

Zdrojový text pro následující příklady

(celý článek)

České noviny:

<https://www.ceskenoviny.cz/zpravy/2496170>

Želivský klášter má nově uspořádanou knihovnu, historický mobiiliář 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. 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.

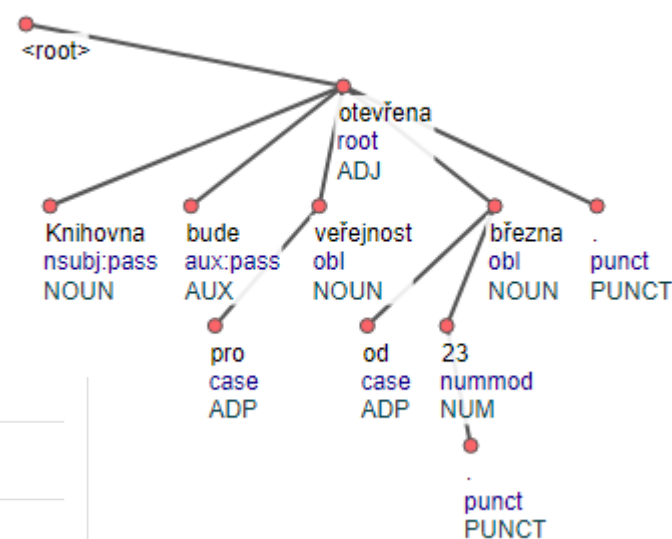
NLP v UDPIpe

- <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

Knihovna bude pro veřejnost otevřena od 23. března .

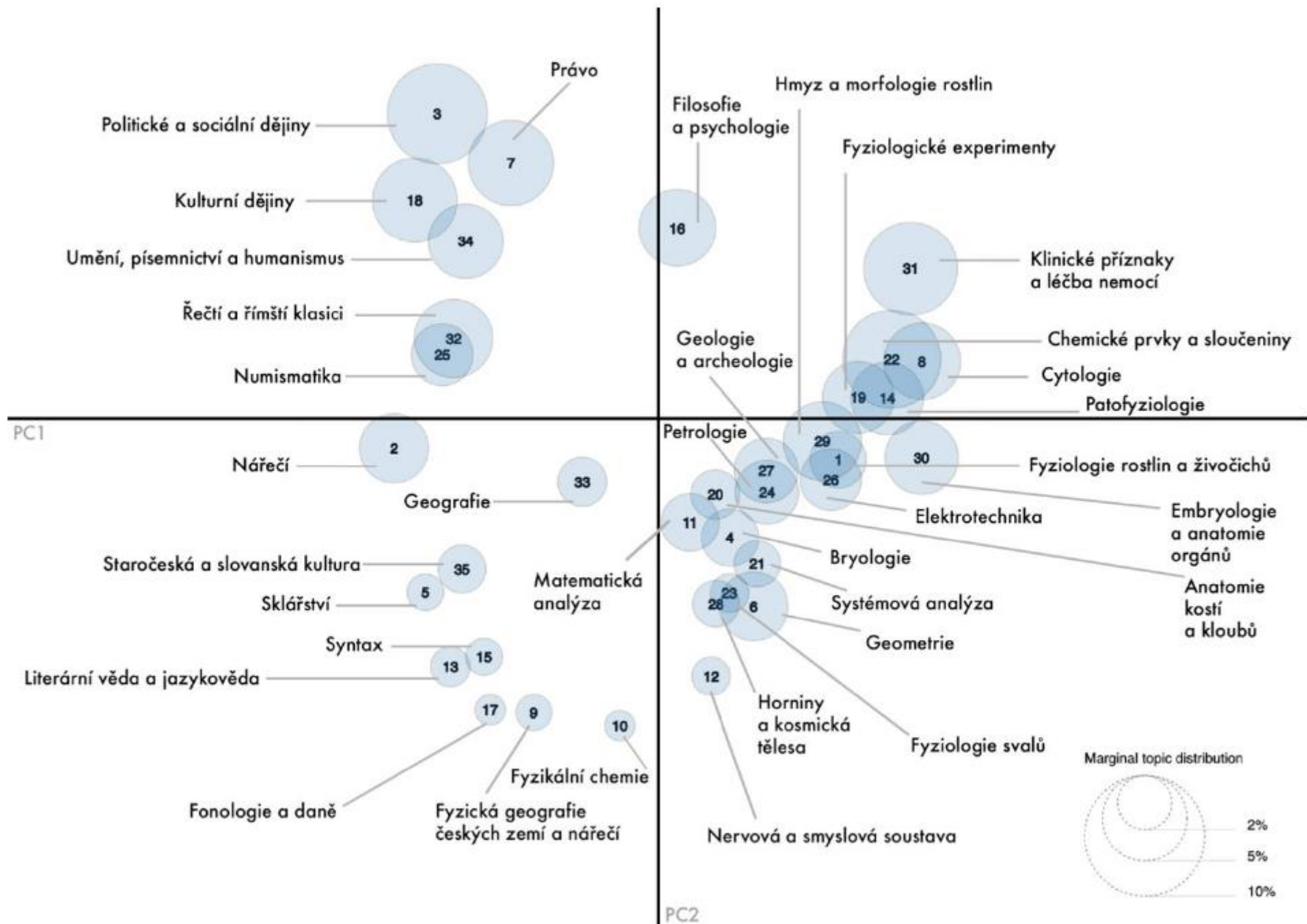


NLP: další možné kroky (>>)

- modelování témat
- analýza sentimentu
- analýza textových sítí: NER (rozpoznání pojmenovaných entit)
- stylometrie
- digitální edice textů
- ...

Modelování témat

- cílem je odhalit skryté tematické struktury v souboru dokumentů
- latentní Dirichletova alokace (LDA)
 - vstup: soubor článků
 - výstup: Téma 1 = "věda, výzkum, univerzita", Téma 2 = "kniha, literatura, autor", ...
- příklad: publikační činnost České akademie věd a umění 1890-1910 ([Filip Kersch 2023](#)) >>>



Analýza sentimentu

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

Analýza sentimentu

Zadání:

Urči sentiment následujícího textu na škále pozitivní - negativní - neutrální: [text článku]

Odpověď modelu Gemma 4 26B A4B:

Sentiment tohoto textu je **neutrální s mírně pozitivním nádechem**.

Zdůvodnění:

Neutrální prvek (informativní charakter): Text je primárně reportážní zprávou, která fakticky informuje o dokončení rekonstrukce, otevření nové knihovny a expozic. Popisuje historická fakta, technické detaily (materiál nábytku, financování z programu IROP) a historický kontext.

Pozitivní prvek (pokrok a obnova): Text zprávu zabarvuje úspěchem – klášter se díky investicím (40 milionů korun) opravuje, knihovna je opět uspořádaná, cenné svazky jsou zabezpečeny a prostor se otevírá veřejnosti. Je to zpráva o "vzkvétání" a obnově památky.

Negativní prvek (historická tragédie): V textu se objevují negativní informace, ale ty jsou použity pouze jako historický kontext (znáčená destrukce nábytku za komunismu, spálení knih, využití budovy jako léčebny). Tyto negativní aspekty slouží k zdůraznění hodnoty současného úspěchu, nikoliv k vyvolání aktuálního negativního sentimentu u čtenáře.

Závěr: Celkový dojem je optimistický zprávou o obnově kulturního dědictví, ale samotný styl psaní zůstává věcný a informativní.

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**

foto

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.

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 hromádách 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 použit 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/>

Měření podobnosti jazyků

- Swadeshův seznam
 - univerzální pojmy
 - všechny jazyky je mají
 - využívané pro lexikostatistiku
 - nejdříve kolem 200 pojmů, později 100, různé verze
 - možné odvodit příbuznost mezi jazyky
- já
- ty
- my
- toto
- tamto
- kdo?
- co?
- ne
- všichni
- mnoho
- jeden
- dva
- ...

Korpusová lingvistika a její nástroje

- 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 slootovorných 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 fóru Poradna ČNK [↗](#) (registrace nutná).

kon textSlovo v kostceSyDMorfioKWordsTreqCalcMapkakorpusdbQuitaUpGramatiKatAkalexalpha

Čeština

▼

stále

×

pořád

×

furt

×

+

Search

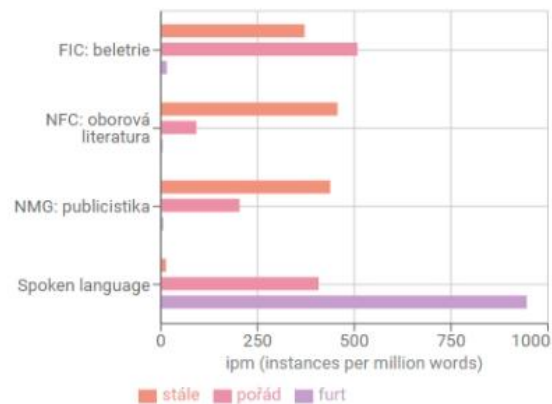
▼ Frequency information

Basic characteristics

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

Source: [syn_v8](#)

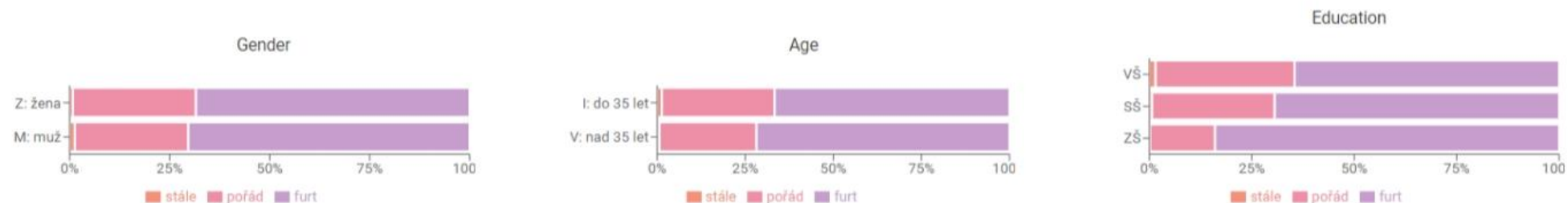
Frequency according to a text type



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

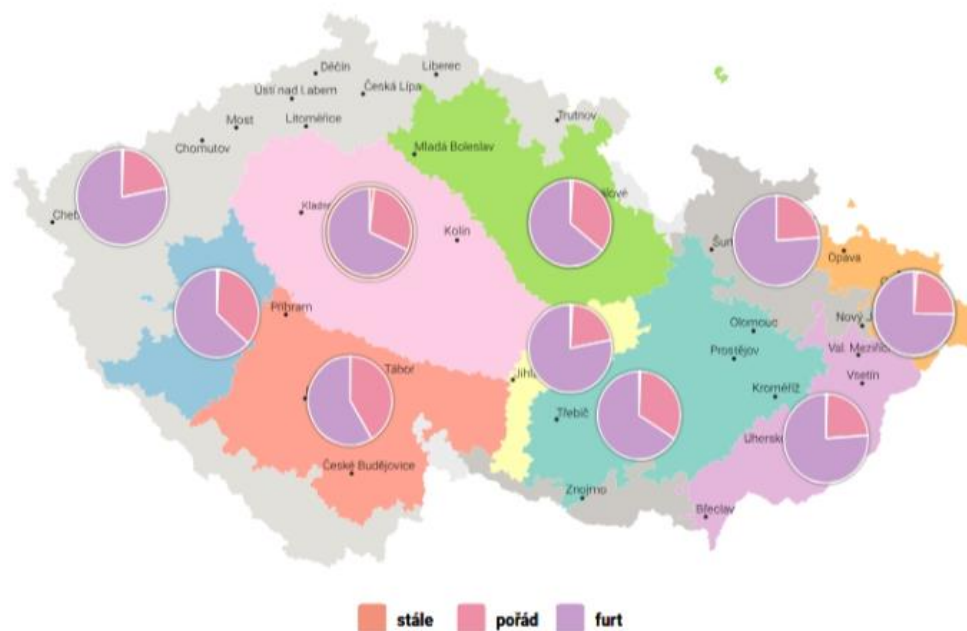


Frequency comparison



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](#)

Další nástroje (příklady)

- Voyant Tools
 - <https://voyant-tools.org/>
- Python: knihovny
 - nltk
 - spaCy
 - gensim
 - stanza
 - ...
 - knihovny pro hluboké a strojové učení

Dnešní cvičení

1. Analýza textu – frekvenční analýza, wordcloud, témata dle LLM
2. Měření podobnosti jazyků – lexikální podobnost, dendrogram

Čemu se věnovat nebudeme:

- extrakce pojmenovaných entit
- tematické modelování (poměrně složitý postup)