Přísudek je v této větě podmět

Jak technologie proměňují práci novinářů - 8. prosince 2021

Barbora Hladká hladka@ufal.mff.cuni.cz

Ústav formální a aplikované lingvistiky MFF UK

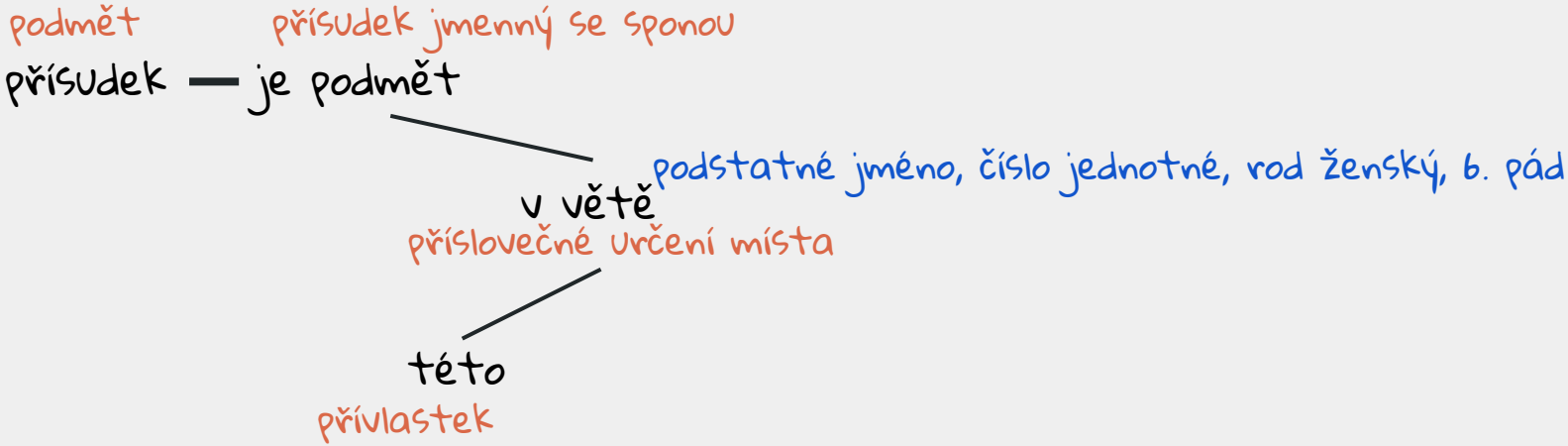
Důležitost lingvistického zpracování textů v jejich vytěžování

- texty, např. novinové články ze serveru iRozhlas

  nebo stenozáznamy z jednání Parlamentu České republiky
- lingvistické zpracování, např.
  - tvaroslovný rozbor
  - větný rozbor
  - rozpoznávání osob, institucí, geografických míst, …
- vytěžování textů, např. rozpoznávání citačních zdrojů

Lingvistické zpracování

Přísudek je v této větě podmět.



Do **Dejvického divadla** jsem vzal vlastně svůj ročník ze školy, z loutkárny **DAMU**. Zůstat v **Praze** a mít vlastní divadlo je pro studenty lákavá nabídka a všem se to tehdy líbilo. Nikdo z okolí mi neměl za zlé, že jsem vzal celý ročník. Bylo jich na začátku devět nebo deset a zůstali tři. Což jsem si i tak představoval, vzpomíná na založení **Dejvického divadla** režisér, scénárista a herec **Miroslav Krobot**, host **Osobnosti Plus**.

Lingvistické zpracování automaticky

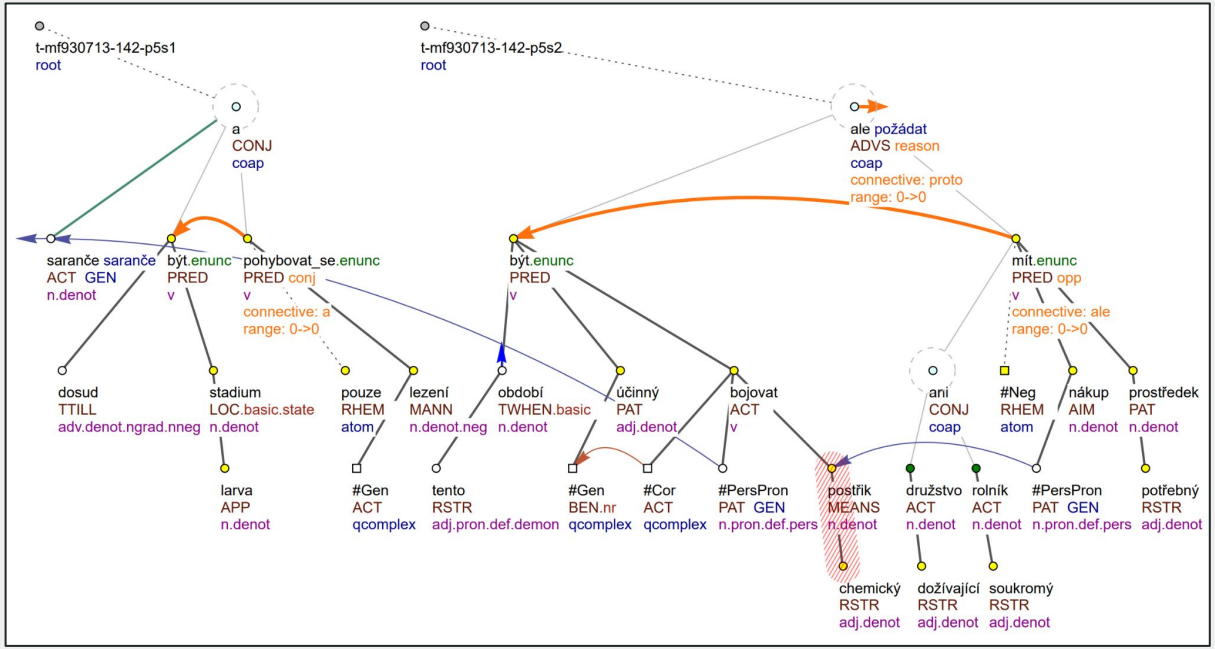
- tvaroslovný a větný rozbor pomocí UDPipe

  https://lindat.cz/services/udpipe/
- rozpoznávání pojmenovaných entit pomocí NameTag

  https://lindat.cz/services/nametag/
- vyzkoušet může kdokoli bez nutnosti další instalace
- nejenom pro češtinu
- = systémy strojového učení naučené na anotovaných korpusech

Anotované korpusy

Rodina Pražských závislostních korpusů (url) + další korpusy pro jiné jazyky



Vytěžování textů :: analýza citačních zdrojů

České ministerstvo zahraničí nemá zatím podle mluvčí Michaely Lagronové …

Že dva roky stará nahrávka Švábenského opravdu souvisí s aktuální kauzou, potvrdily serveru iROZHLAS.cz dva důvěryhodné zdroje blízké vyšetřování.

Sociolog a zakladatel platformy PAQ research Daniel Prokop nicméně upozorňuje, že příjmová chudoba není nejlepší ukazatel stavu společnosti.

Petříček ve středu uvedl, že na kandidátce nebude kvůli tomu, že tím byla narušena její demokratická tvorba.

Analýza citačních zdrojů

- rozpoznat fráze, které odkazují ke zdrojům

České ministerstvo zahraničí nemá zatím podle mluvčí Michaely Lagronové …

Že dva roky stará nahrávka Švábenského opravdu souvisí s aktuální kauzou, potvrdily serveru iROZHLAS.cz dva důvěryhodné zdroje blízké vyšetřování.

Sociolog a zakladatel platformy PAQ research Daniel Prokop nicméně upozorňuje, že příjmová chudoba není nejlepší ukazatel stavu společnosti.

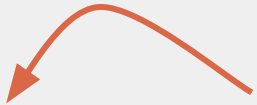
Petříček ve středu uvedl, že na kandidátce nebude kvůli tomu, že tím byla narušena její demokratická tvorba.

Ruční anotace citačních zdrojů a frází

- Domácí úkol pro 230 posluchačů kurzu Digitální komunikace a práce s informacemi na FSV UK, 1.12.2021-7.1.2022
  https://ufal.mff.cuni.cz/anotace-citacnich-frazi-v-datech-irozhlas

V článcích (cca 2 200) ze serveru iRozhlas

- označit a klasifikovat citační zdroje
- označit citační fráze
- spojit zdroje s frázemi



O víkendu by mohla vlna teplého počasí ještě vygradovat, napsala agentura AFP.
oficiální, nepolitický

Klasifikace zdrojů v anotacích Kredit: Václav Moravec

- Nepojmenované
  - Anonymní
  - Anonymní částečně
- Pojmenované
  - Oficiální - institucionální příslušnost
    - politický
    - nepolitický
  - Neoficiální - „obyčejní lidé“

K čemu potřebujeme ruční anotace

- hledáme vzorce citačních frází

ukázka frází označených studenty

315	podle
215	Podle
188	řekl
179	uvedl
177	říká
78	uvedla
67	vysvětluje
66	dodává
60	dodal
55	popisuje
52	tvrdí
41	řekla

fráze v základních tvarech = vzorce

530	podle
270	uvést
173	říkat
90	dodat
74	informovat
68	oznámit
62	vysvětlovat
66	dodávat
55	tvrdit
64	popisovat
39	upozorňovat
38	psát
36	popsat

K čemu potřebujeme ruční anotace

- ze vzorců umíme automaticky generovat citační fráze,
  tj. skloňujeme a časujeme

Příklad: lemma = upřesnit

Předseda poslaneckého klubu Jan Chvojka neupřesnil, zda ČSSD podpoří na předsedu sněmovny Radka Vondráčka (ANO).

"Chceme, aby nejrizikovější skupina divoce žijících zvířat byla chráněná," uvedla ministryně Matečná. Upřesnila, že zákaz bude platit i ...

V podobném duchu se neslo vyjádření Vladimíra Šibora z útvaru zvláštních činností policie, který zpracovává informace od operátorů pro celou policii. Upřesnil, že

K čemu potřebujeme ruční anotace

- podle vzorců tvoříme pravidla
- např. podle koho čeho (osoba)

```
[lemma="podle"] ([upos="ADJ|NOUN"])* <name_type = "PER"> []* [xpos="....2........."] []*
</name_type> within s
```

za předložkou podle je řetězec podstatných nebo přídavných jmen ukončený vlastním jménem ve 2. pádě

Příklad: Oblast je podle místopředsedy Asociace cestovních kanceláří Jana Papeže pro turisty nezajímavá a málo navštěvovaná .

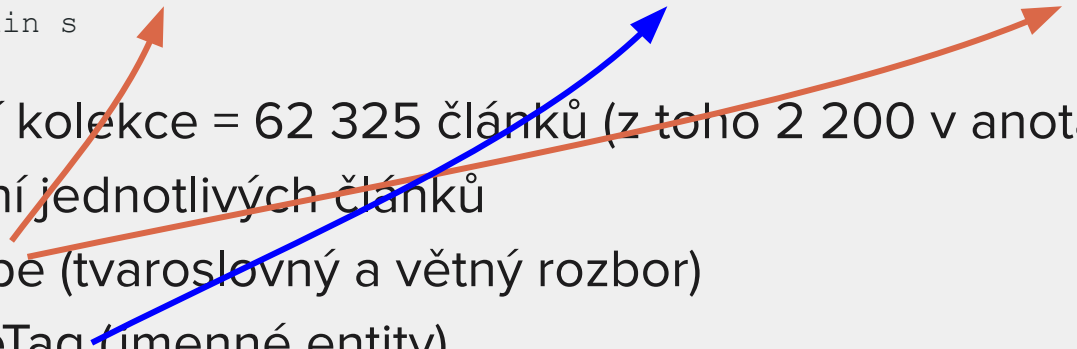
Aplikace pravidel

```
[lemma="podle"] ([upos="ADJ|NOUN"])* <name_type = "PER"> []* [xpos="....2........."] []*
</name_type> within s
```

- kompletní kolekce = 62 325 článků (z toho 2 200 v anotační úloze)
- zpracování jednotlivých článků
  - UDPipe (tvaroslovný a větný rozbor)
  - NameTag (jmenné entity)

Aplikace pravidel

```
[lemma="podle"] ([upos="ADJ|NOUN"])* <name_type = "PER"> []* [xpos="....2........."] []*
</name_type> within s
```



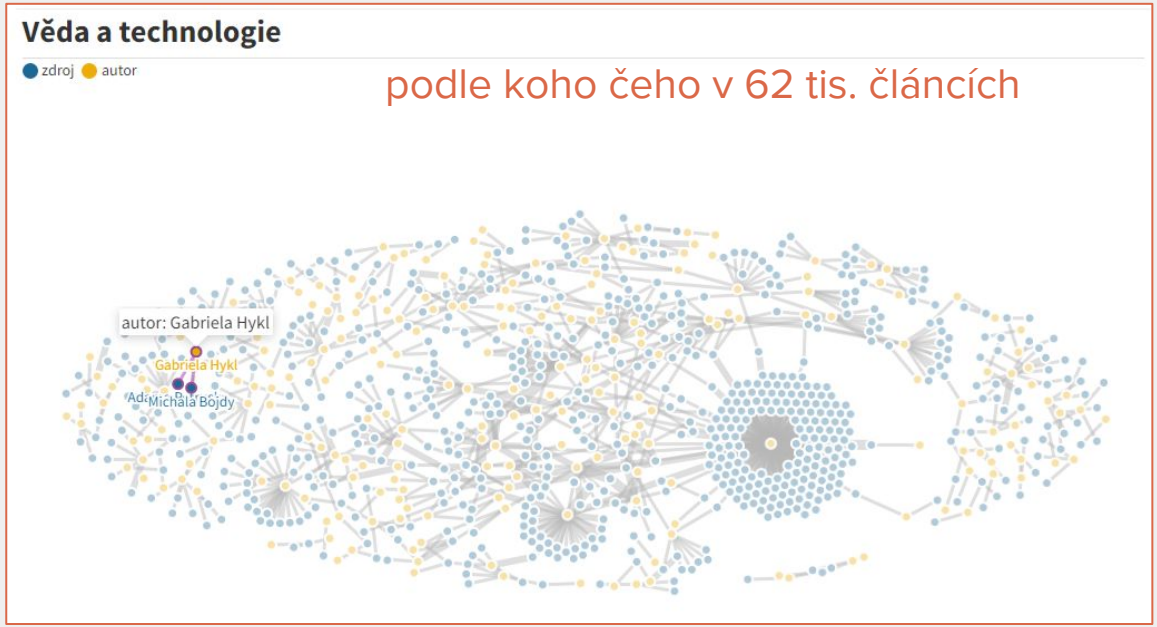
- kompletní kolekce = 62 325 článků (z toho 2 200 v anotační úloze)
- zpracování jednotlivých článků
  - UDPipe (tvaroslovný a větný rozbor)
  - NameTag (jmenné entity)
  - pravidla (nebo strojové učení?)

K čemu potřebujeme ruční anotace

- klasifikované citační zdroje jako příklady pro strojové učení

Analýza citačních zdrojů

- až skončí anotace, až vytvoříme a otestujeme pravidla, až natrénujeme modely
- vizualizace



Lépe!

ParCzech (url, projekt LINDAT/CLARIAH-CZ)

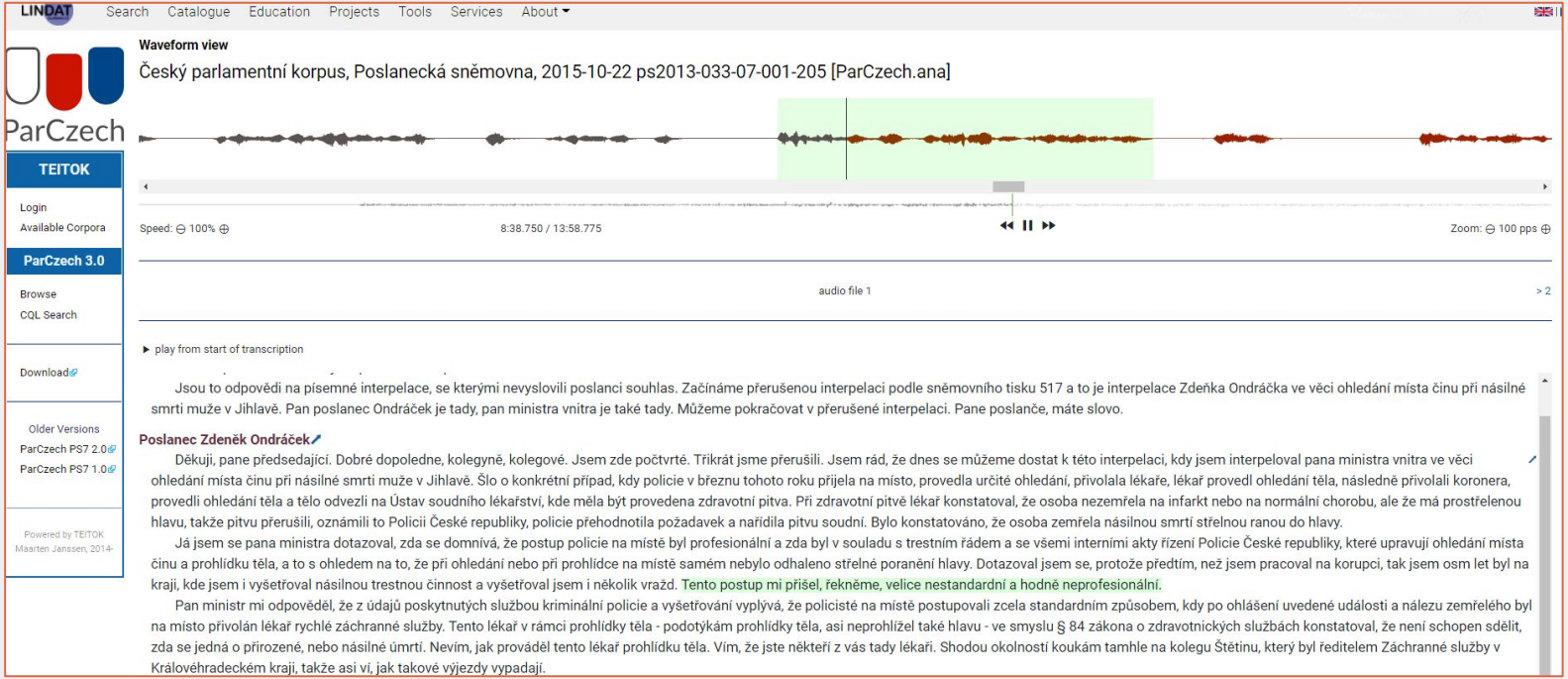
- kompilace stenozáznamů z jednání PČR do anotovaného korpusu

ParCzech a iRozhlas

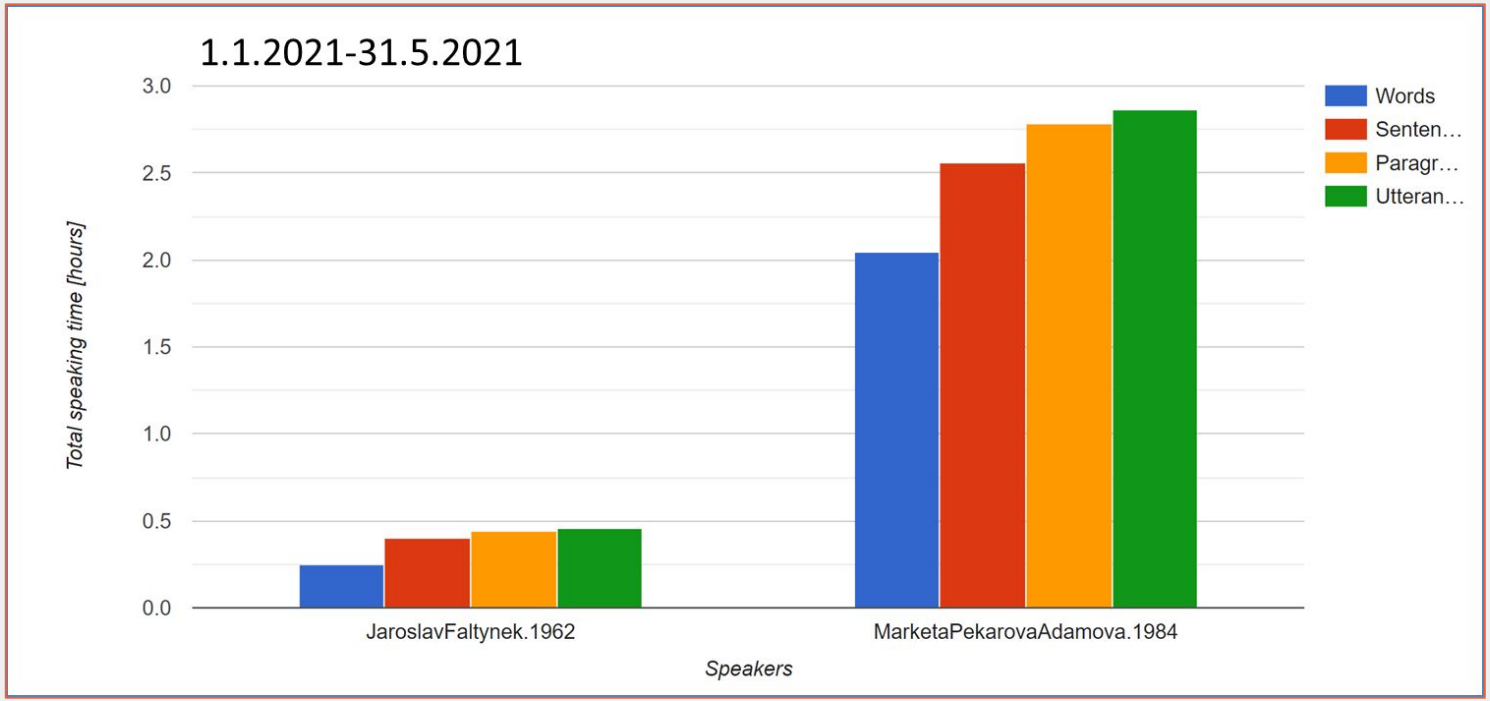
- konverze do stejného formátu (TEI)
- UDPipe
- NameTag
- automatická detekce a klasifikace citačních zdrojů
  - pravidla a strojové učení z iRozhlasu

ParCzech :: délky promluv poslanců

- audio zarovnané s textem



ParCzech :: délky promluv poslanců



Poděkování

- Děkuji Matyášovi Koppovi a Jiřímu Mírovskému z ÚFAL MFF UK za skvělou spolupráci.
- Děkuji studentům FSV UK za anotace.
- Analýza citační zdrojů probíhá v rámci projektu TL05000057 Signál a šum v éře Žurnalistiky 5.0 - komparativní perspektiva novinářských žánrů automatizovaných obsahů.
- Projekt ParCzech je podporován výzkumnou infrastrukturou LINDAT/CLARIAH-CZ.