Projekt HIČKOK: podíl ÚJČ

Jiří Pergler

Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka

pergler@ujc.cas.cz

Historický text a jeho interpretace 2024, Kozojedy, 23.–25. října 2024

Termíny, personální obsazení, úvazky

- od září 2023 do listopadu 2026
- řešitelský tým: Jiří Pergler (spoluřešitel za ÚJČ), Anna Michalcová, Olga Navrátilová, Jana Zdeňková, Ondřej Svoboda

- **září 2023 – únor 2024**
  - v ÚJČ celkem 1 úvazek

Jiří Pergler	0,2
Anna Michalcová	0,2
Olga Navrátilová	0,2
Jana Zdeňková	0,2
Ondřej Svoboda	0,2

- **březen 2024 – listopad 2026**
  - v ÚJČ celkem 2 úvazky

Jiří Pergler	0,5
Anna Michalcová	0,45
Olga Navrátilová	0,6
Jana Zdeňková	0,4
Ondřej Svoboda	0,05

Úkoly ÚJČ

- zodpovědnost za část monitorovacího korpusu pokrývající období od nejstarších textů do 18. století (tj. stará a střední čeština)

- 2 hlavní úkoly:

  1) **poskytnout texty pro korpus**
     - edice textů, které jsou podkladem pro STB a SDTB
     - z velké části technická práce, většina v prvním půlroce

  2) **vytvořit nástroj na jejich automatickou lemmatizaci a morfologické značkování**
     - ve spolupráci s ÚFAL MFF UK (Dan Zeman)
     - hlavní část práce na projektu, časově nejnáročnější

Texty pro monitorovací korpus

- elektronické edice textů, které vznikly dříve v OVJ ÚJČ
  - 430 textů (296 stč., 134 střdč. (dle rozdělení pro STB/SDTB))
  - cca 8,3 mil. tokenů (cca 7,1 mil. stč., cca 1,2 mil. střdč.)

1) **technická příprava**
   - cílem je zajištění kompatibility napříč korpusem
   - úprava formátu souborů, uvádění metadat apod. (Ondřej Svoboda)

2) **klasifikace podle žánrů / textových typů**
   - korpus pokrývá celé období psané češtiny, není možná jednotná klasifikace žánrů
   - pro stč. a střdč. období pouze dva typy: *fikce*, *ne-fikce* (i to velmi problematické)
   - fikce (157): např. rytířské texty, lyrika, písně, satiry, modlitby, biblické texty, pasionály, ...
   - ne-fikce (273): např. korespondence, právní texty, lékařské texty, cestopisy, kroniky, homilie, ...

Texty pro monitorovací korpus

3) **rozdělení textů na „jádro“ a „doplněk“**

- 2 vzájemně rozporné cíle: vyváženost (chronologická, žánrová, ...) vs. maximální velikost korpusu
  - řešení: vyvážené jádro, maximální jádro+doplněk
- situace ve stč. a střdč. je specifická
  - není možné dosáhnout vyváženosti jako v následujících obdobích
  - existence variantních rukopisů -> v korpusu více variant téhož textu
- -> v jádru pouze jedna varianta textu, ostatní v doplňku
- kritéria pro zařazení do jádra: kompletnost, stáří, chybovost (+ zkratky k ESSČ)
  - často jdou proti sobě

Lemmatizace a morfologické značkování

- 2 otázky:
  1) jak mají vypadat?
  2) jak je vytvořit?

- **jak má vypadat lemmatizace?**
  - pro starší texty různé možnosti, např. lemma odpovídající počátku období (např. StčS/ESSČ k roku 1300), konci období, současnému jazyku, konkrétnímu výskytu v korpusu, prvnímu/poslednímu doloženému výskytu lexému
  - možnost rozlišení lemma vs. hyperlemma
  - cílem je jednotná lemmatizace pro celý korpus
  - **<u>řešení</u>**:
    - lemmata formálně odpovídají současné češtině
    - včetně nepravidelných hláskových změn (*usta* > *ústa*, *anjel* > *anděl*, *poprslek* > *paprsek*, *mhla* > *mlha*)
    - u slov nedochovaných do současné češtiny se rekonstruuje nč. hlásková podoba (*jěšutenstvie* > *ješitenství*)
    - přechody mezi flexivními typy (*uvrci* > *uvrhnout*, *jmě* > *jméno*, *hřěbí* > *hřebík*, *cizý* > *cizí*)

Jak má vypadat morfologické značkování?

- jednotné pro celý korpus
- standard **Universal Dependencies** (UD, https://universaldependencies.org/)
  - „a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages“
  - výrazný český podíl (Dan Zeman, Jan Hajič, ...)
  - přes 160 jazyků, všechny relevantní slovanské jazyky (vč. staroslověnštiny, staré východní slovanštiny)
  - pro češtinu několik synchronních korpusů, včetně ČNK (Intercorp v16ud)
- anotace slovních druhů (POS), morfologických rysů (features), nikoli syntaktické vztahy (zatím)
- anotační pravidla z korpusů současné češtiny, úpravy kvůli specifikům starší češtiny
  - doplnění kategorií: aorist sigmatický/asigmatický, imperfektum, supinum
  - rozšíření kategorií: např. duál, pád u jmenných tvarů adjektiv, stupeň tamtéž
  - mnoho dalších jevů k řešení: např. interpretace klitik *-ť*, *-ž*, interpretace kondicionálového auxiliáru, interpretace jednotlivých typů zájmen, interpretace spojek, částic a dalších funkčních slov (např. *ještě*), interpretace předložky *v* > *u* atd. atd.

Jak korpus lemmatizovat a označkovat?

- různé přístupy, srov. STB: seznam hyperlemmat ze stč. slovníků, k nim generovány všechny tvary podle vzorů, popisů alternací a hláskových změn
  - výhody: důkladnost, spolehlivost
  - nevýhody: časová náročnost, není desambiguace
- zvolený přístup: automatická lemmatizace a značkování pomocí nástroje **UDPipe**
  - počítačový nástroj pracující s natrénovanými jazykovými modely
  - https://lindat.mff.cuni.cz/services/udpipe/
  - výsledek rovnou desambiguovaný
  - k dispozici jazykové modely pro současnou češtinu
- je třeba vytvořit jazykové modely pro starou češtinu, střední češtinu a češtinu 19. století (pro češtinu 20. století se využije novočeský model)
  - etalonové korpusy (ručně lemmatizované a označkované), na kterých se nástroj natrénuje
  - úkol pro ÚJČ: 2 etalonové korpusy: staročeský a středněčeský

Etalonové korpusy

- velikost korpusů
  - dostatečně vysoká, aby se nástroj zvládl kvalitně natrénovat
  - dostatečně nízká, abychom je zvládli v daném čase kvalitně vytvořit
  - 100 000 slov / korpus
- 2 úkoly:
  1) volba textů
  2) lemmatizace a značkování

Volba textů pro etalonové korpusy

- kritérium: reprezentativnost
  - etalony by měly ideálně obsahovat všechny jazykové jevy / typy textů, s nimiž se lze v textech daného období setkat
  - poměr těchto jevů by měl ideálně odpovídat jejich skutečnému zastoupení v textech, které mají být automaticky lemmatizovány a značkovány
- texty vybírány z dostupných el. edic: podmnožina monitorovacího korpusu
- pouze úryvky textů, ideální velikost cca 2000 slov
- vyváženost chronologická
  - datace pramene vs. památky – obě řešení problematická, vychází se z datace pramene (datace památky v metadatech není)
    - preference textů, kde mezi tím není velká časová vzdálenost
      - -> mj. preference starších variantních pramenů
  - u obou korpusů 4 časová období (stč. 50letá, střdč. 75letá) s rovnoměrným zastoupením
    - ve stč. málo textů z 1. pol. 14. stol., proto se s 14. stol. pracuje jako s „dvojobdobím“

Volba textů pro etalonové korpusy

- vyváženost „žánrová“ (fikce vs. ne-fikce)
  - v každém období rovnoměrně zastoupeny oba typy (pokud je to možné)
- počty slov:

Etalon	100 000
1 období	25 000
1 textový typ (fikce / ne-fikce) v období	12 500

- tj. při ideálním úryvku o 2000 slovech vychází na 1 textový typ v 1 období cca 6 textů
  - některé texty kratší než 2000 slov, někde naopak méně než 6 textů -> delší úryvky
- pro každé období + textový typ vybrány texty
  - často nebylo moc na výběr
- následně z každého textu vybrán úryvek
  - u textů kratších než 2000 slov typicky celý text
  - u delších textů cca třetina úryvků ze začátku textu, třetina z prostřední části, třetina z konce textu

Složení staročeského etalonu

14. stol.	fikce	AlxH (1533), AlxBM (1742), BiblDrážď (2061), BiblDrážď (2042), HradMagd (2025), HradProk (2075), HradSat (2250), KristA (2118), MastMuz (2033), ModlKunh (639), PasMuzA (2014), UmučRajhr (1811), ŽaltU (839), ŽaltWittb (2088)
	ne-fikce	DalV (7626), PřípJiř (25), PříslFlaš (1495), RadaOtcR (1375), ŘádKorA (7345), ŠtítKlem (7577)
1. pol. 15. stol.	fikce	AlexPovD (2094), Astar (2084), BiblOl (2131), BiblOl (2059), OtcB (2139), PodkU (2132)
	ne-fikce	HusKorD_35 (1272), LékFrantMuz (2038), LyraMat (1395), MajCarA (2052), MartKronA (2041), PrávŠvábC (2030), ZrcSpasK (2091)
2. pol. 15. stol.	fikce	AsenM (2119), BawArn (2104), BiblKladr (2380), PovOl (2176), Svár (2083), TkadlB (2113)
	ne-fikce	BřezSnářM (2126), CestMandM (2203), HusSvatokup (2170), LékŽen (2097), LetMuz (2129), PrávJihlA (2158)

Složení středněčeského etalonu

1501–1575	fikce	MendlEzop (2170), FrantPráv (3173), PetříkPov1554 (721), PiccoloSen1516 (1969), PísKnobloch (1262), SatŠvorc1538 (3213)
	ne-fikce	Artik1554 (2156), Kuch (2134), ListDorotaZKrumlova1 (2094), PeresteriusKáz (2143), VespAmer (2038), VočehMor (2149)
1576–1650	fikce	DačProstopravda (3228), KronJovAnte1586 (3299), PirckChlouba1597 (3297), ŠubarPís1612 (3160)
	ne-fikce	ListŽofieAlbínkaZHelfenburku (2132), NovDiv1612 (2098), RosslinZahr1588 (2066), ŠubarKáz (2159), VelPředmlApatéka (2184), ŽalManž1615 (2182)
1651–1725	fikce	Enšpígl1700 (4137), EzopŽivot1696 (3981), KomKancTisk (3966), RosaRozml1672 (588)
	ne-fikce	CestPalestina1701 (2812), KnihaSmolBojk (2966), KnihaSmolBojk (2924), VetterIslandia (4036)
1726–1800	fikce	PísNepom1789 (301), PísNepom1799 (354), Vendelín1753 (7731)
	ne-fikce	KnihaSmolJimr (2555), KuchKníž1763 (2789), Lucid1750 (2882), PranNová1766 (2837), PranSedl1766 (2898), ŠolcPam (2755)

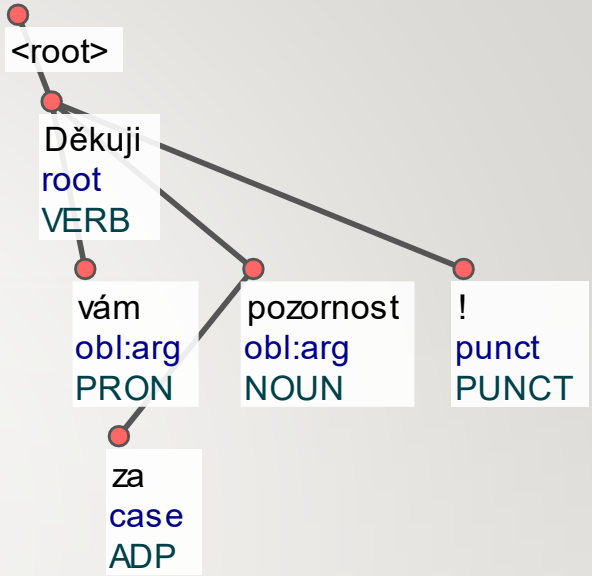
Lemmatizace & značkování etalonových korpusů

- časově nejnáročnější (od března 2024)
- poloautomaticky: UDPipe s novočeským modelem generuje preanotaci, následně manuální kontrola/oprava
  - data ve formátu .tsv, práce v tabulkovém procesoru (Excel)
  - v preanotaci celkově většina lemmat i značek správně, zároveň mnoho zásadních chyb
  - prvních 10 opravených textů použito k natrénování nového modelu -> preanotace by měla být kvalitnější
- každý text anotován nezávisle dvěma anotátory
  - následně generován výpis mezianotátorských rozdílů, podle něj vytvořena definitivní verze
    - rozdíly: zřejmé chyby/přehlédnutí, různé interpretace, nejasnost v tom, jak daný jev anotovat
- průběžné konzultace, evidence problematických jevů, diskuse o jejich řešení

Výhled

- **cíle:**
  - provedení anotace všech textových úryvků do etalonových korpusů
  - průběžně přetrénovávání modelů pro kvalitnější preanotaci
  - průběžně řešení problematických jevů, doplňování dokumentace
  - po dokončení průřezová kontrola, finalizace etalonových korpusů
  - následně natrénování staročeského a středněčeského modelu
  - automatická lemmatizace a značkování staročeské a středněčeské části monitorovacího korpusu

Děkuji vám za pozornost!



Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Misc
1	Děkuji	děkovat	VERB	VB-S---1P-AA--1	Mood=Ind|Number=Sing|Person=1|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act	0	root	TokenRange=0:6
2	vám	ty	PRON	PP-P3--2-------	Case=Dat|Number=Plur|Person=2|PronType=Prs	1	obl:arg	TokenRange=7:10
3	za	za	ADP	RR--4----------	AdpType=Prep|Case=Acc	4	case	TokenRange=11:13
4	pozornost	pozornost	NOUN	NNFS4-----A----	Case=Acc|Gender=Fem|Number=Sing|Polarity=Pos	1	obl:arg	SpaceAfter=No|TokenRange=14:23
5	!	!	PUNCT	Z:----------	_	1	punct	SpaceAfter=No|TokenRange=23:24