ČESKÝ NÁRODNÍ
KORPUS

ČESKÝ NÁRODNÍ KORPUS

Historie češtiny v korpusovém kontinuu (HiČKoK)

https://korpus.cz/hickok/

HIČKOK

Historie češtiny v korpusovém kontinuu

OBSAH

1. V RÁMCI ČEHO, KDY a s KÝM
2. CÍLE PROJEKTU
3. JAKÉ ZDROJE MÁME PRO ČEŠTINU K DISPOZICI
4. S JAKÝM MATERIÁLEM REÁLNĚ PRACUJEME
5. CO JE NUTNÉ UDĚLAT
6. JAKÝM ČELÍME VÝZVÁM
7. ZÁVĚR

V RÁMCI ČEHO, KDY A S KÝM

1. veřejná soutěž programu SIGMA - Dílčí cíl 3: Podpora inovačního potenciálu společenských věd, humanitních věd a umění (SHUV)

- doba trvání:
  září 2023 – listopad 2026

- zapojené instituce:
  UK (ÚČNK, ÚFAL); AV ČR (ÚJČ – Oddělení vývoje jazyka) a NK ČR

CÍLE PROJEKTU

a) monitorovací korpus (MKČ) pokrývající všechny etapy vývoje češtiny (13.–21. století)

b) jazykové modely ve schématu Universal Dependencies (UD), které umožní automatickou lingvistickou anotaci textů z libovolného období (připravuje DZ)

c) aplikace Timeline Maker umožňující zkoumat diachronní fenomény v MKČ (celé připravuje VC)

d) online kurz pro studenty a badatele pracující s historickými texty, který poskytne průpravu pro využití vytvořených výstupů

JAKÉ ZDROJE MÁME PRO ČEŠTINU K DISPOZICI

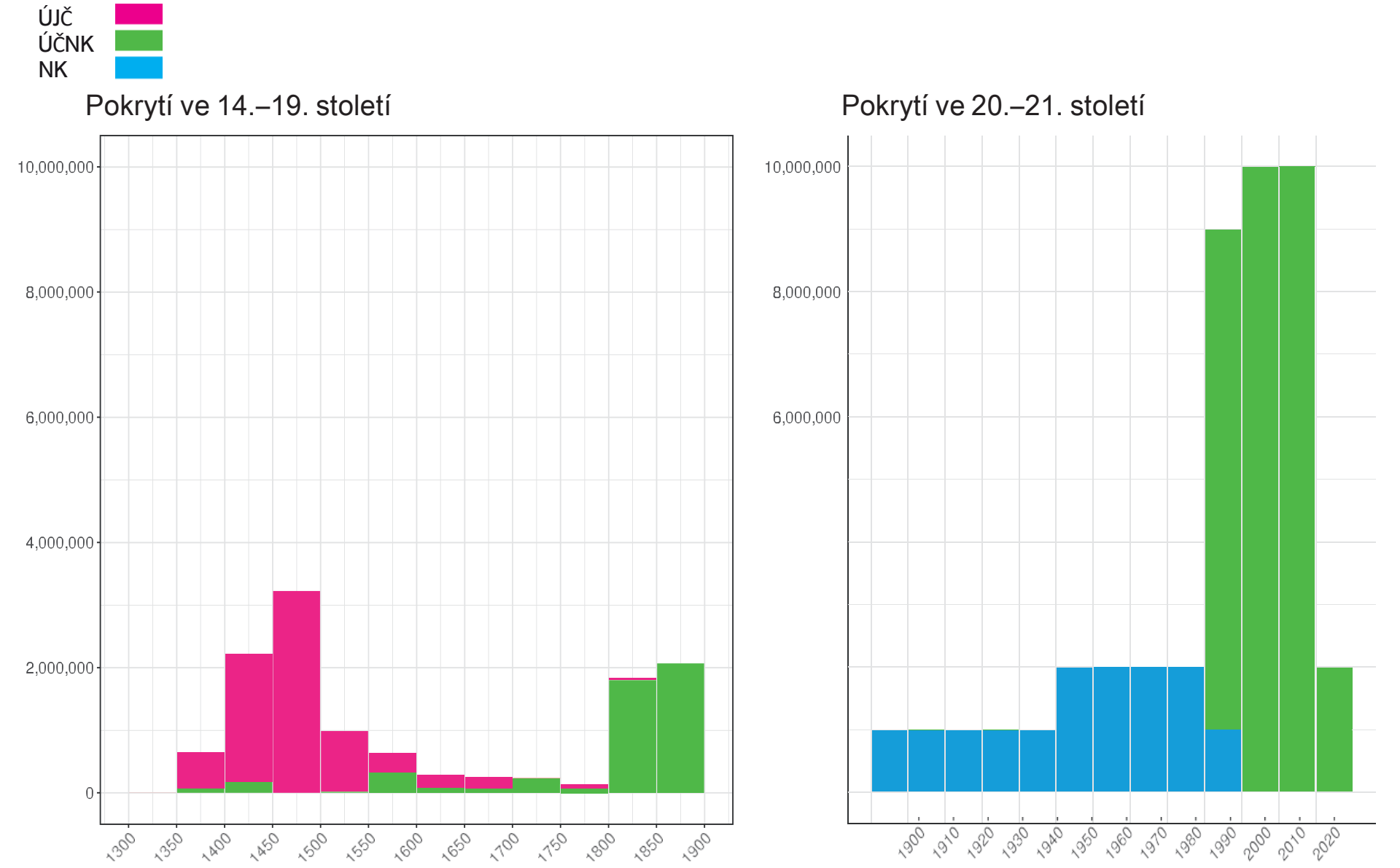
- starší období x období po roce 1990
- po roce 1990 relativně neomezené množství žánrově bohatého textového materiálu v elektronické podobě, a tedy korpusově zpracovatelného
- starší vývojové období:

- poměrně rozsáhlý textový materiál
- deponován v rámci různých knihoven a institucí
- jen zlomek existuje v elektronické (korpusově zpracovatelné) podobě
- texty jsou postupně zpracovávány do elektronické podoby
  - v závislosti na době vzniku textů méně nebo více (časově) náročné > do roku 1842 povětšinou ruční transkripce; po roce 1842 možnost OCR (často velmi nespolehlivé)
- souvislé česky psané texty doloženy od konce 13. století (během historického vývoje v různém množství a žánrové bohatosti)
- zejména texty z 20. stol. jsou ještě zatíženy autorskými právy

S JAKÝ MATERIÁLEM REÁLNĚ PRACUJEME

Monitorovací korpus		Odhad
Období do roku 1800	Tokeny	8,6 mil.
	Texty	429
	Složení	nevyvážené
Období 19. století	Tokeny	3,9 mil.
	Texty	270
	Složení	částečně vyvážené (BEL, ODB, PUB)
Období 1901–1990	Tokeny	13,3 mil.
	Texty	650
	Složení	vyvážené (BEL, ODB, PUB)
Období po r. 1991	Tokeny	31 mil.
	Texty	1550
	Složení	vyvážené (BEL, ODB, PUB)
Tokeny celkem		**56,8 mil.**
Texty celkem		**2899**

S JAKÝM MATERIÁLEM REÁLNĚ PRACUJEME

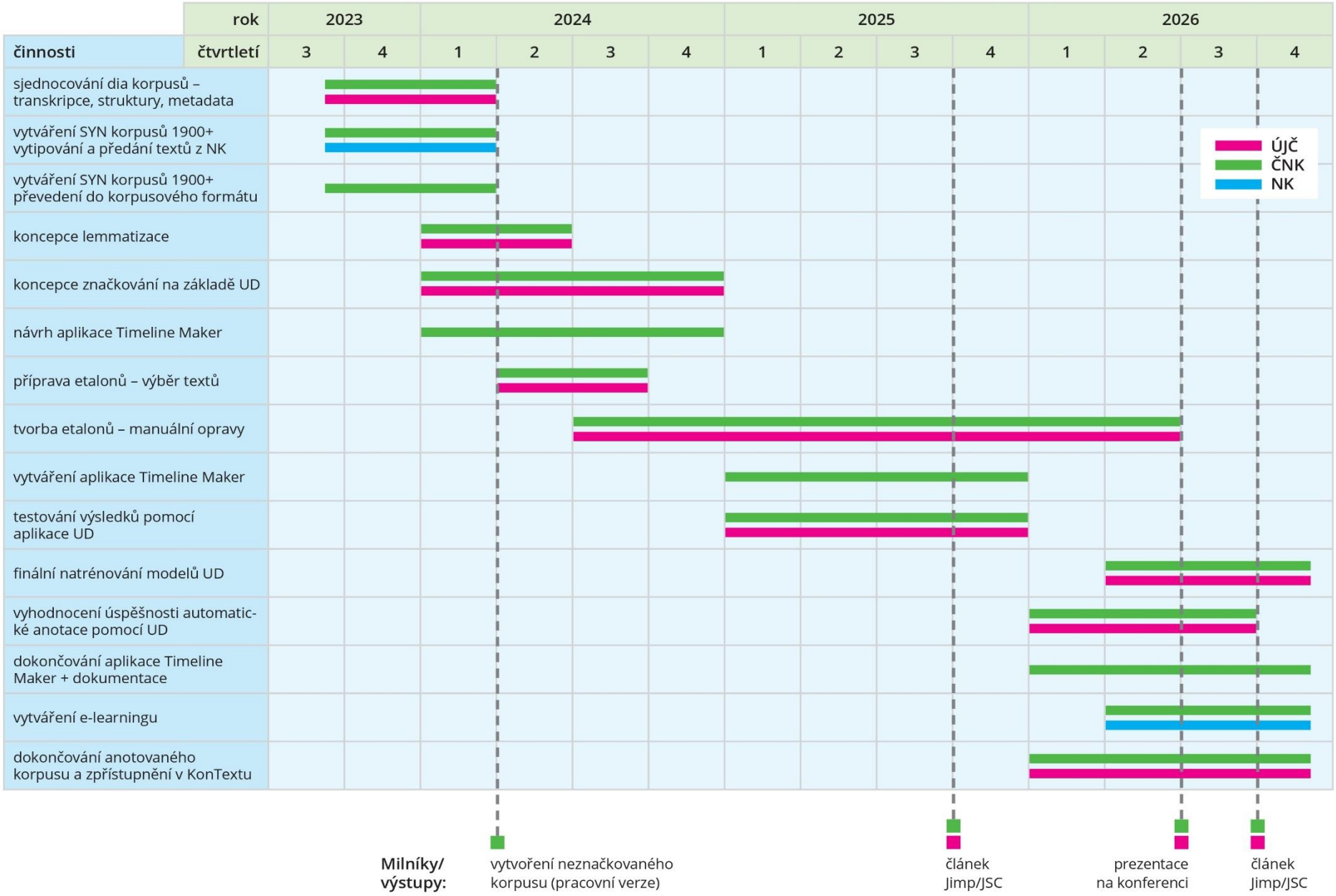


CO JE NUTNÉ UDĚLAT

...aby vznikl oanotovaný MKČ podle nejnovějšího standardu UD a abychom získali modely ve schématu UD:

- vytvořit etalony (ručně značkované trénovací korpusy), pokud možno co nejvíce žánrově a časově vyvážené:
- 3 stotisícové etalony pro období
  - 1300–1500; 1500–1800 a 19. století
- časové úseky pro výběr textů do etalonů
  - 1300–1500: 50letá období (ale 100letý úsek pro 14. stol. > 1. pol. 14. stol. – málo relevantních dat)
  - 1500–1800: 75letá období
  - 1800–1900: 10letá období
  - 20. století: nevytváří se speciální etalon

CO JE NUTNÉ UDĚLAT



JAKÝM ČELÍME VÝZVÁM

…VYPLÝVAJÍ ZE ZPRACOVÁNÍ DAT Z RŮZNÝCH OBDOBÍ, JAK Z HLEDISKA DOBY VZNIKU TEXTŮ, TAK Z HLEDISKA DOBY PRVNÍHO ELEKTRONICKÉHO ZPRACOVÁNÍ

- **lemmatizace** > pro projekt nutné vytvořit jednotící přístup, ať už je zpracování textů v různých institucích jakékoli (přístupy jsou různé a mimo projekt i různé zůstanou)
- **žánrová specifikace** > snaha o řešení smysluplné pro celé zpracovávané kontinuum
  (obligatorní 3 žánry: BEL, ODB a PUB)
  - lze uplatnit pouze cca v období 1800–současnost
  - pro období 1300–1800 obligatorní jen 2 žánry > BEL a ODB, resp. cokoli nebeletristického (pro jiné řešení by bylo nutné nejprve provést žánrovou analýzu staré a starší češtiny)

JAKÝM ČELÍME VÝZVÁM

žánrová vyváženost > dosáhnout vyváženosti je v různých obdobích různě náročné

13.–18. stol.:

- užší výběr textů
- problematika vzniku pramene a pozdějších opisů
- jiná žánrová situace

19. stol.:

- lze realizovat navrženou koncepci
- s ohledem na texty, které jsou k dispozici, není řešení mnoho

20. stol. (do r. 1990):

- zdálo by se, že s žánrovou architekturou nebudou potíže
- ale výběr textů není neomezený, smysluplná vyváženost vyžaduje kreativitu

JAKÝM ČELÍME VÝZVÁM

- Popis standardu: https://universaldependencies.org
- Tagger UDPipe: https://lindat.mff.cuni.cz/services/udpipe/

- primárně morfologická anotace textů ze staré a střední češtiny (etalonů) > za využití novočeské UD a formou oprav a „učení“

- konverze morfologicky anotovaného etalonu češtiny 19. století do formátu UD (anotace 19. stol. vzniká v ÚČNK vlastním na míru vytvořeným postupem)
- morfologická anotace textů z 20. a 21. stol. pomocí novočeské UD

> syntéza = model UD morfologicky anotující češtinu v celém jejím vývoji

HiČKoK – ZÁVĚR

POZITIVA A JEDINEČNOSTI PROJEKTU

- získání korpusově zpracovaných jazykových dat, která reflektují celý vývoj češtiny v rámci jednoho korpusu a jsou jednotně anotována pomocí nejnovějších standardů UD (unikátní nejen v ČR, ale i celosvětově – diachronní korpusy Helsinského typu na starou angličtinu aj. zahrnují jen část jazykového vývoje a nejsou dotaženy do současnosti)
- vznik jazykových modelů v nejnovějším standardu UD, které bude možné využít pro anotaci jakéhokoli jiného korpusu
- díky spolupráci s NK ČR získání jazykových dat z 20. století
- aplikace Timeline Maker zprostředkuje nový rozměr v rámci diachronního zkoumání dat
- možnost rozšíření povědomí o výsledcích projektu pomocí e-learningu

Díky za pozornost!

Projekt HiČKoK (č. TQ01000072) byl na období realizace na období 09/2023 – 11/2026 podpořen TA ČR v rámci programu na podporu aplikovaného výzkumu a inovací SIGMA