Staročeská (hyper)lemmatizace

Vznik příspěvku byl podpořen projektem Ministerstva školství, mládeže a tělovýchovy č. LM2015081 „Výzkumná infrastruktura pro diachronní bohemistiku“ (akronym RIDICS) v rámci Projektu velkých infrastruktur pro VaVal.

RIDICS

Ondřej Svoboda, ÚJČ AV ČR, svoboda@ujc.cas.cz

Staročeská (hyper)lemmatizace a (nejen) morfologické značkování

I. určení: staročeská textová banka (primárně)
II. možnosti stč. lemmatizace/značkování
III. jak vzniká stč. morfologická databáze
    - historie + generování tvarů
IV. ukázky

Ve třech krocích

I. generování slovních tvarů (od všech hyperlemmat)
II. morfologická databáze (word, hyperlemma, tag, is_lemma, variant_tag)
III. (víceznačně a částečně) anotovaný korpus

Pojmy

- „kuoním“ ⇒ lemma „kuoň“ (fonologicky odpovídá tvaru)
- hyperlemma „kóň“ (odpovídá fonologickému systému češtiny kolem r. 1300, tak jako staročeské slovníky: GbSlov, StčS, MSS, ESSČ)
- hyperlemma je zároveň jedním z lemmat: „kóň“, „kuoň“, „kůň“

Určení: stč. textová banka

- vzniká v oddělení vývoje jazyka od roku 2006
- korpus (ručně) transkribovaných českých textů cca. do roku 1500
- transkripce standardizuje pravopis (včetně interpunkce), ale zachovává fonologicky relevantní rysy jazyka
- dostupný ve Vokabuláři webovém a nově i prostřednictvím KonTextu
- více než 200 pramenů různého rozsahu, téměř 5,5 milionu tokenů (asi 200 tisíc typů)

Možnosti (automatické) anotace

- (baseline) moderní data a nástroje
- „postaršení“ novoč. morfologické databáze, „ponovočeštění“ stč. textu (Hana et al., 2012)
- lemmata pro 2000 nejčastějších tvarů (ibid.)
- formální popis stč. lexika (Jínová/Synková et al., 2014–)
- morf. databáze (word, (hyper)lemma, tag) a vedlejší aplikace
- (v budoucnu) značky pro varianty/mutace (à la NovaMorf) a trénovací desambiguovaný korpus

Stč. morfologická databáze

- Pavlína Synková: formální popis stč. apelativní deklinace (disertační práce, 2017):
  - Gebauerova Hist. mluvnice jazyka českého + ověřování v interní textové bance, rukopisech a jejich edicicích
  - vzory (kmenotvorné přípony + koncovky) podle (psl.) kmene a rodu
  - seznam (hyper)lemmat (ze slovníků): vzor, alternace tvarotvorného základu, omezení paradigmatu

Stč. morfologická databáze

- Pavlína Synková: formální popis stč. apelativní deklinace (disertační práce, 2017):
  - vzory (kmenotvorné přípony + koncovky) podle (psl.) kmene a rodu
  - seznam (hyper)lemmat: vzor, alternace tvarotvorného základu, omezení paradigmatu
  - hláskové a pomocné/formální změny
  - systém + výjimky
  - vzory ve webové aplikaci

Funkce generátoru

I. (hyper)lemma hadačka, vzor žena, alternace

EO bezkoncovkove_tvary K_1eK_2 tvary_s_koncovkou K_1K_2

II. odtržení koncovky Nsg -a ⇒ tvarotvorný základ <hadačk>

III. koncovce -a odpovídá alternace K_1K_2

IV. "hadačk" = "(hada)K_1K_2" ⇒ K_1=č K_2=k

V. přiřazení alternací TZ ke koncovkám paradigmatu:

-ø = hadaček-, ostatní hadačk-

VI. spojení tvarotvorných základů a koncovek a aplikace „hláskových“ změn: Lsg *hadačk-ě > hadačcě > hadačce

Vývoj lemmatizace

2017	2018
* apelativní substantiva (číslo a pád)	+ neohebné slovní druhy, příslovce a 6. slovesná třída
* hyperlemmata (bez desambiguace)	+ lemmata (bez desambiguace)
* atributivní (brněnské) morfologické značky	+ poziční (pražské, Hajičovy) značky à la korpusy SYN*
* aplikace stč. slovní tvary a KonText interně v „alfaverzi“	⇒ obě aplikace na webu (s manuály)

Výhled na rok 2019 a dál

- práce na slovesech (vzory pracěvati (substráty), býti, negace, 5. a 4. třída – prošu), později na adjektivech, zájmenech, číslovkách, propriích
- anotace agregátů (nejdříve -s/-ť/-ž)
- webové rozhraní k vývoji českého hláskosloví (s možností aplikace změn)

- samostatné variantní/ mutační tagy (NovaMorf) – *kuoním: globalMutationUOflexiveVariantIEMsoundChangeÍ
- desambiguace: asistovaná / pravidlová / ML

Ukázky

- kózle, dobrě, na, peskovati
- slovesa = více morfémů/částí: tematický sufix, prefixy (negace), flexe participií…
- zdrojová data: vzory (2), parametry lemmat, hláskové změny

Skloňování

Lemma	Vzor	Omezení
mládě	subst.n.nt-kmen.kuře	žádné (plné paradigma)

	SG	DU	PL

NOM	• ~~mláďě~~
	  ◦ mládě
		• ~~mláďěte~~
		  ◦ mláděte
			• mláďata

GEN	• ~~mláďěte~~
	  ◦ mláděte
		• mláďatú
		  ◦ mláďatou
			• mláďat

DAT	• ~~mláďěti~~
	  ◦ mláděti
		• mláďatma
		• mláďatoma
			• mláďatóm
			  ◦ mláďatuom
			  ▪ mláďatům
			  ◦ mláďatom

ACC	• ~~mláďě~~
	  ◦ mládě
		• ~~mláďěte~~
		  ◦ mláděte
			• mláďata

VOC	• ~~mláďě~~
	  ◦ mládě
		• ~~mláďěte~~
		  ◦ mláděte
			• mláďata

LOC	• ~~mláďěti~~
	  ◦ mláděti
		• mláďatú
		  ◦ mláďatou
			• ~~mláďětech~~
			  ◦ mládětech

Časování

kupovati | verb.6.kupovati | žádné (plné paradigma) | sl

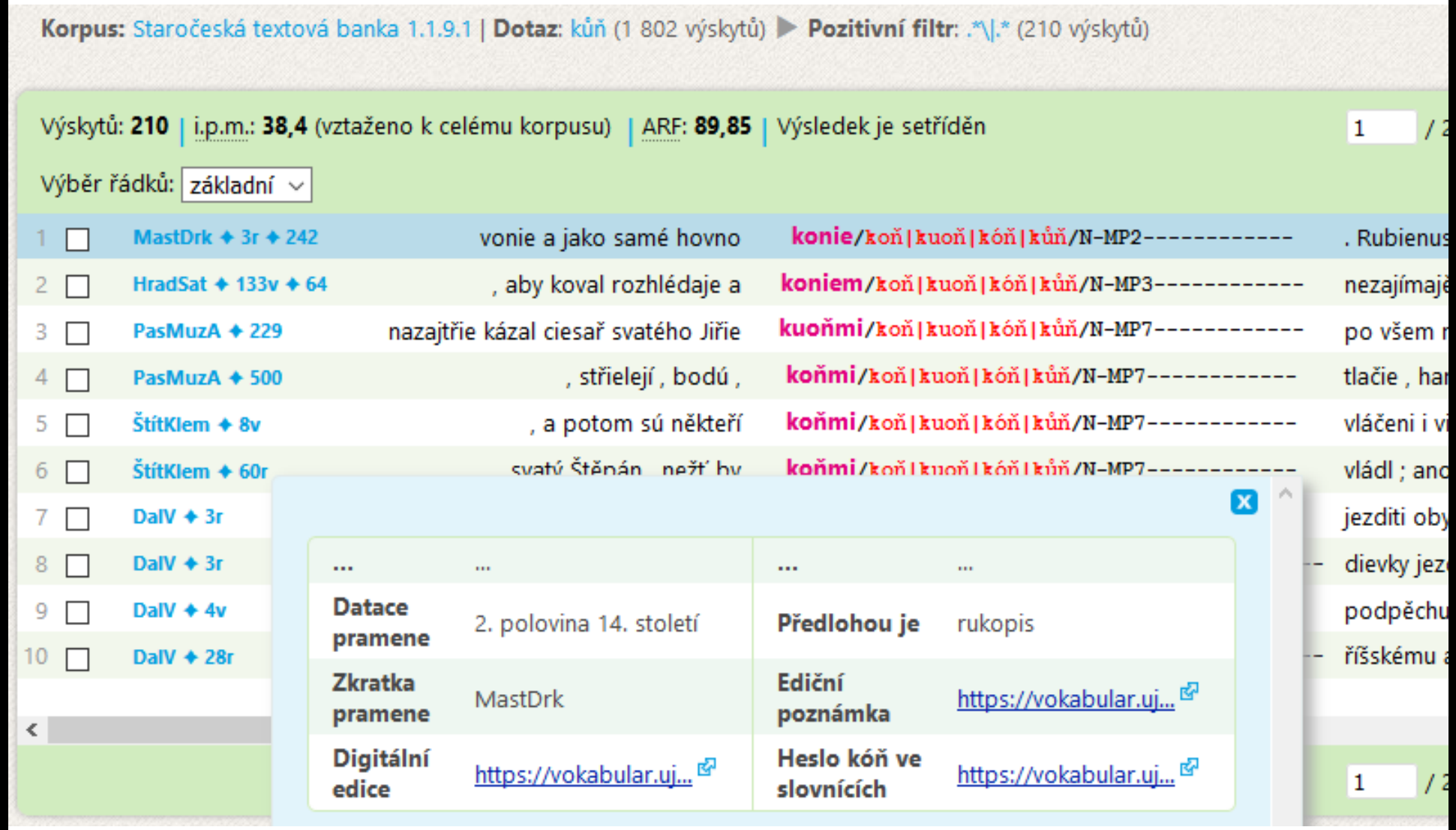
participle

nt

	SG	DU	PL
NOM	• kupujě ◦ kupuje • kupující ◦ kupující	• kupujúce ◦ kupujíce • kupujúc ◦ kupujíc	• kupujúce ◦ kupujíce • kupujúc ◦ kupujíc

slovní tvar	kupujíc
tvarotvorný základ	kup
zakončení	-ujíc (PAR.NT.DU.NOM.m)
hlásková změna	jú > jí (u-i-fronting, 2. až 3. čtvrtina 14. století)
kde	na konci tvarotvorného základu (na posledním grafému + možná i v zakončení)
identifikátor	kupujíc (PAR.NT.DU.NOM)
odvozeno z	kupujúc (PAR.NT.DU.NOM)
(atributivní)	k5gMnDc1alm S

Upravený KonText



Hláskové změny

- cítit > cejtit, vozík > vozejk, …
- Kí > Kej; K = c, z, …
- od konce 14. století a ve století 15. (cca 1386–1500)

Vybraná bibliografie

- Hana, Lehečka, Feldman, Černá, Oliva: Building a Corpus of Old Czech. LREC 2012, pp. 9–15
- Jínová/Synková, Lehečka, Oliva: Describing Old Czech Declension Patterns for Automatic Text Analysis. Mundo Eslavo 2014, pp. 7–17
- Synková: Popis staročeské apelativní deklinace (...). FF UK, disertační práce
- Synková, Lehečka, Svoboda: Na cestě k lemmatizaci staročeských textů: data, software, aplikace. SALI 2018, pp. 66–84