Implementing an Old Czech word forms generator

RASLAN, Karlova Studánka, 2019

This work was fully supported by RIDICS ("Research Infrastructure for Diachronic Czech Studies"), a project LM2015081 funded by the Czech Ministry of Education, Youth and Sports of the Czech Republic.

RIDICS

Ondřej Svoboda, Czech Language Institute of the CAS

Purpose and origins

⇒ lemmatization and tagging of “Old Czech text bank”, a 5.5M corpus hosted at Vokabulář webový, “Web Vocabulary”

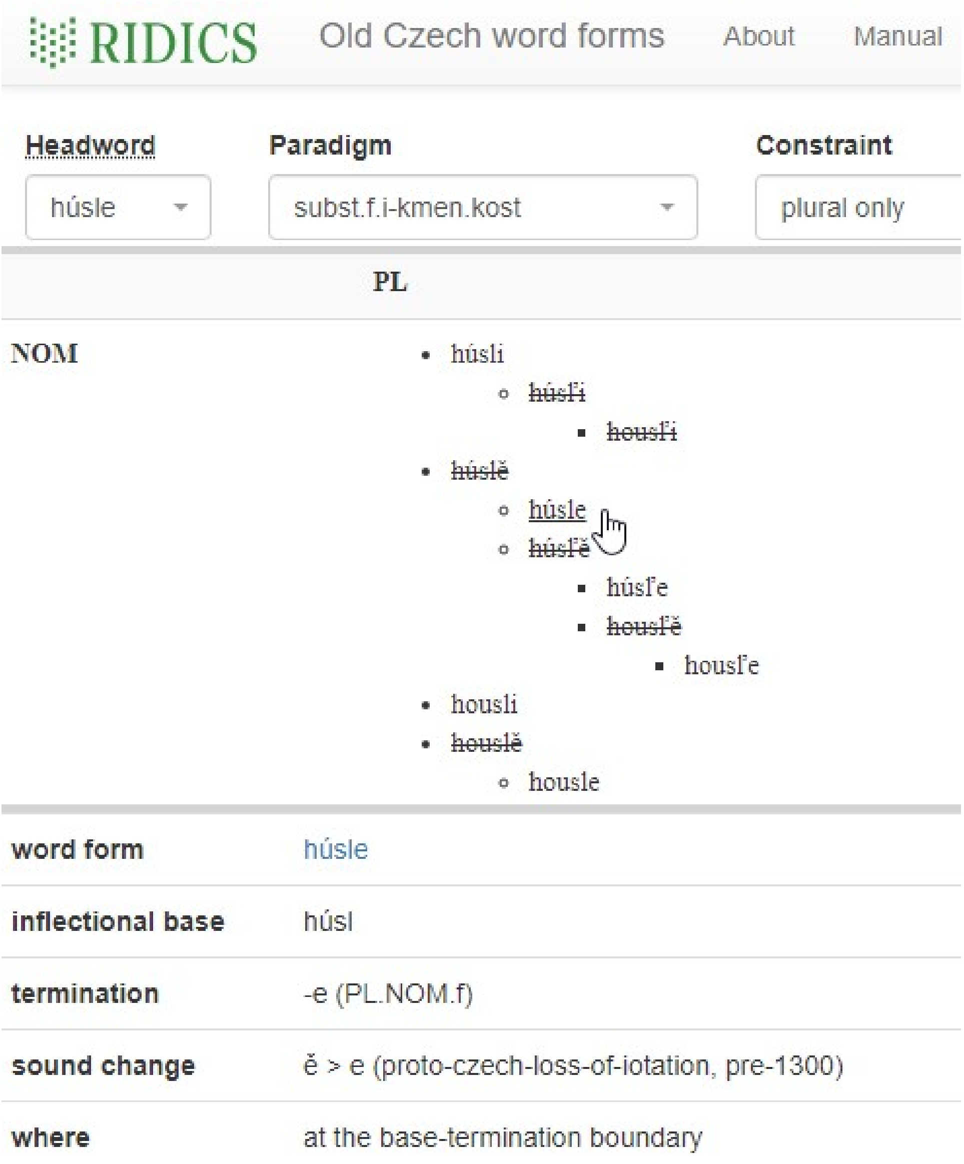
- formal description of Old Czech (OC) common nouns in Pavlína Synková’s Ph.D. thesis (2017)
- compact, generative approach

⇒ headwords + alternations of inflectional bases + declension paradigms + sound/formal changes

A glimpse at foundations

headwords (& alternations)	paradigms	sound changes
húsle ("fiddle"), plural-only	noun.kost (declension)	"proper": ú > ou cě > ce
⋮		⋮
pracovati ("to work"), imperfective	verb.kupovati (conjugation)	"auxiliary": ~~*lě~~ = le ~~*kě~~ = ce
⋮		⋮
krátcě ("briefly")	adv.mocně (gradation)	
↑ 1300's rendering / reconstruction ↑		until cca 1500

A teaser (declension)



Headword: húsle

Paradigm: subst.f.i-kmen.kost

Constraint: plural only

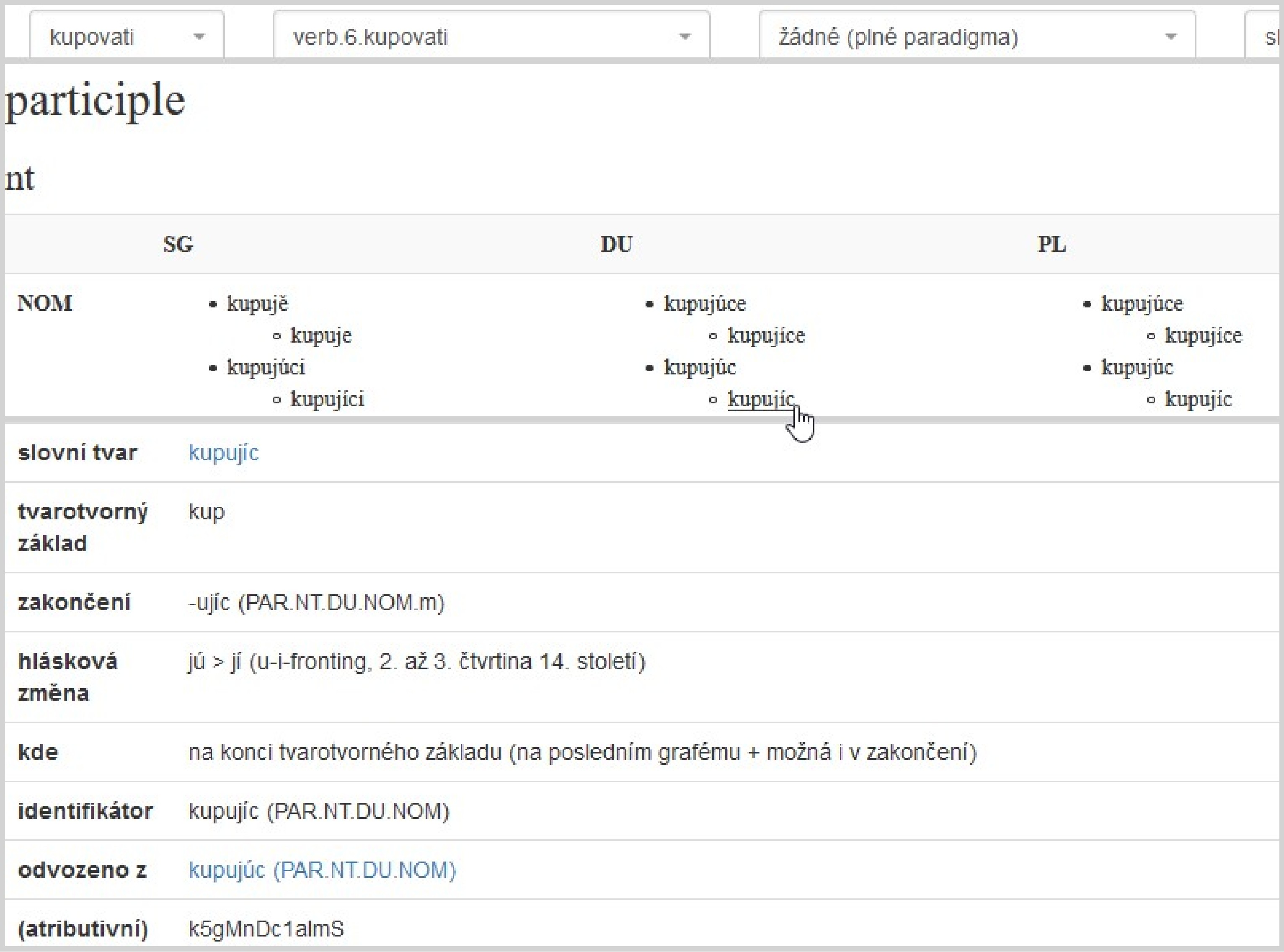
PL

NOM

- húsli
    - ~~húsľi~~
        - ~~housľi~~
- ~~húslě~~
    - húsle
    - ~~húsľě~~
        - húsľe
        - ~~housľě~~
            - housľe
- housli
- ~~houslě~~
    - housle

word form	húsle
inflectional base	húsl
termination	-e (PL.NOM.f)
sound change	ě > e (proto-czech-loss-of-iotation, pre-1300)
where	at the base-termination boundary

Another teaser: conjugation



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

participle

nt

	SG	DU	PL
NOM	• kupujě ∘ kupuje • kupujúci ∘ kupujíci	• 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)	k5gMnDc1aImS

The process (for given headwords)

	phase	component	example
1.	assign paradigm	core	húsle: noun.kost
2.	accomodate to headword (constraint)	paradigm module	húsle: plural
3.	“develop” terminations	sound change engine	K’ě > K’e ě > (l)-e
4.	“stemming” ⇒ raw base	core	húsle – (l)-e = húsl- mládě – ě = mlád-
5.	“develop” the raw base ⇒ inflectional base(s)	sound change engine	mlád- = mláď-

The process (for given headwords)

	phase	component	example
6.	assign terminations to bases	core	okn- + -o, . . . oken- + -∅
7.	develop bases	sound change engine	obal- > vobal- + -ova-ti
8.	generate negative bases	core	ne + obal-, ne + vobal-
9.	“early” word forms	core	moř- + -ě, húsl- + -ě
10.	develop word forms	sound change engine	mořě > moře *húslě = húsle

Paradigms module

Cases ⇒ prepositions

```xml
<cases>
  <nominative case="NOM/1" base="true"/>
  <genitive case="GEN/2"/>
  <dative case="DAT/3"/>
  <accusative case="ACC/4"/>
  <vocative case="VOC/5"/>
  <local case="LOC/6"/>
  <instrumental case="INS/7"/>
</cases>
```

Numbers and (reused) cases ⇒ nouns

```xml
<declension>
  <singular number="SG/S" likeParadigm="cases" base="true"/>
  <dual number="DU/D" likeParadigm="cases"/>
  <plural number="PL/P" likeParadigm="cases"/>
</declension>
```

```xml
<conjugation>
  <persons>
    <firstPerson person="1/1"/>
    <secondPerson person="2/2"/>
    <thirdPerson person="3/3"/>
  </persons>
  <present mood="IND/I/indicative" type="PI/I">
    <singular number="SG/S" likeParadigm="persons"/>
    <dual number="DU/D" likeParadigm="persons"/>
    <plural number="PL/P" likeParadigm="persons"/>
  </present>
  <future type="FUT/B" likeParadigm="present"/>
  <imperative type="IPV/R">
    <singular number="SG/S">
      <secondPerson person="2/2"/>
      <thirdPerson person="3/3"/>
    </singular>
    <dual number="DU/D" likeParadigm="persons"/>
    <plural number="PL/P" likeParadigm="persons"/>
  </imperative>
  <imperfect type="IPF/M" likeParadigm="present"/>
  <aorist type="AOR/O">
    <sigmatic subtype="S" likeParadigm="present"/>
    <asigmatic subtype="NS" likeParadigm="present"/>
  </aorist>
```

```xml
<participle type="PAR">
  <nt subtype="NT" likeParadigm="declension" type="/S"/>
  <s subtype="S" type="/D">
    <singular number="SG/S"/>
    <dual number="DU/D"/>
    <plural number="PL/P"/>
  </s>
  <l subtype="L" likeParadigm="s" type="/A"/>
  <n subtype="N" likeParadigm="declension" type="/N"/>
  <t subtype="T" likeParadigm="declension" type="/N"/>
</participle>

<gradation>
  <positive grade="POS/1" base="true"/>
  <comparative grade="CMP/2"/>
  <superlative grade="SUP/3" useForm="CMP" prepend="naj"/>
</gradation>

<pos>
  <noun paradigm="declension"/>
  <verb paradigm="conjugation"/>
  <adverb paradigm="gradation"/>
  <preposition paradigm="cases" government="true"/>
  <conjunction paradigm="none"/>
  <particle paradigm="none"/>
  <interjection paradigm="none"/>
</pos>
```

Affixes

- superlatives come with a prefix (naj-), verb participles carry two suffixes; adjectives will combine negation and gradation

```
<supine>
  <termination>
    <stemSuffix>ova</stemSuffix>
    <ending>t</ending>
  </termination>
</supine>
<participle>
  <nt>
    <singular>
      <nominative>
        <termination homonymy="gender-m+n" gender="m">
          <stemSuffix>uj</stemSuffix>
          <participleSuffix>0</participleSuffix>
          <ending>ě</ending>
        </termination>
```

Sound changes & future work

- multi-alternation verbs (4th class – prošu), and later adjectives, pronouns, numerals
- ability for sound changes to target prefixes (in verbs, deverbatives, gradables)
- Old Czech phonology live presentation

- standalone variant/ mutation tags (NovaMorf) – *kuoním: globalMutationUO flexiveVariantIEM soundChangeÍ
- disambiguation: semi-automatic, rule-based, ML

Selected bibliography

- Synková: Description of Old Czech common nouns' declension (...). Faculty of Arts, Charles University, Prague. 2017. Ph.D. thesis
- Synková, Lehečka, Svoboda: Towards lemmatization of Old Czech text: data, software, applications. SALI 2018, pp. 66–84

Thank you for your attention!