IZDELAVA IN ANALIZA DIGITALIZIRANE ZBIRKE PAREMIOLOŠKIH ENOT

Saša Babič, ZRC SAZU, Inštitut za slovensko narodopisje, sasa.babic@zrc-sazu.si
Tomaž Erjavec, Institut »Jožef Stefan«, Odsek za tehnologije znanja, tomaz.erjavec@ijs.si

TRADICIONALNE PAREMIOLOŠKE ENOTE V DIALOGU S SODOBNO RABO

OBJAVA ZBIRKE NA CLARIN.SI

Zbirka je objavljena pod odprto licenco CC-BY na repozitoriju CLARIN.SI: Babič, Saša et al. 2022. Collection of Slovenian paremiological units Pregovori 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1455.
Zbirka je dostopna tudi na konkordančnikih noSketch Engine in KonText CLARIN.SI

PRIPRAVA GRADIVA

Ročni pregled in urejanje paremioloških enot v dveh Excelovih tabelah: v prvi 59.543 paremioloških enot in v drugi 2.742 virov.

Združitev Excelovih tabel v OpenRefine za poenotenje korektorskih opomb in kategorij označevanja pregovorov.

Prenos podatkov v delovno bazo SQLite za enostavnejše lektorske popravke ter zaznavo uporabljenih črkovnih naborov (dajnčica, metelčica, bohoričica, gajica).

Izvoz podatkov v TSV format: po ročnem in strojnem urejanju gradivo vsebuje 36.349 relativno urejenih paremioloških enot in 2.515 virov:

- Razpredelnica z viri vsebuje za vsak bibliografski vir njegov identifikator, identifikator z izvornega seznama virov, zaporedno številko vira (ki tudi združuje vire, ki spadajo v nadrejeno enoto), letnico izida (in letnico prvega izida, kjer se ta razlikuje), ime vira (avtor, naslov) ter kategorizacijo vira v 18 kategorij, npr. Leposlovje in literarjenje, Muzejske zbirke, Periodika – pratike in koledarji, Ustni viri itd.
- Razpredelnica s paremiološkimi enotami vsebuje identifikator enote, zaporedno številko iz izvornega seznama enot, seznam identifikatorjev virov skupaj s številko strani, na kateri je enota v viru omenjena, diplomatično transkripcijo.

ZAPIS TEI

Celotna zbirka je bila formirana kot en TEI dokument (element <TEI>) s kolofonom (element <teiHeader>) in besedilnim delom (<text>).

```
<ab xml:id="PREG-00-00001" n="1">
    <seg xml:lang="sl-bohoric" xml:id="PREG-00-00001.dipl" type="dipl">Bres muje
        fe zhreul ne obuje.</seg>
    <seg xml:lang="sl" xml:id="PREG-00-00001.crit" type="crit">Brez muje se čreul
        ne obuje.</seg>
    <bibl corresp="#bib14.1">
        <biblScope unit="page">202</biblScope>
    </bibl>
    <bibl corresp="#bib23.1">
        <biblScope unit="page">51</biblScope>
    </bibl>
    <bibl corresp="#bib7.1">
        <biblScope unit="page">524</biblScope>
    </bibl>
</ab>
```

```
<seg xml:id="P1.crit.norm.ana" type="crit.norm.ana" xml:lang="sl">
 <s xml:id="P1.crit.s1">
   <w ana="mte:Sg" msd="UPosTag=ADP|Case=Gen" lemma="brez" xml:id="P1.crit.s1.t1">Brez</w>
   <w ana="mte:Ncfsg" msd="UPosTag=NOUN|Case=Gen|Gender=Fem|Number=Sing" lemma="muja" xml:id="P1.crit.s1.t2">muje</w>
   <w ana="mte:Px------y" msd="UPosTag=PRON|PronType=Prs|Reflex=Yes|Variant=Short" lemma="se" xml:id="P1.crit.s1.t3">se</w>
   <w norm="čevlj" ana="mte:Ncmsn" msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing" lemma="čevlj" xml:id="P1.crit.s1.t4">čreul</w>
   <w ana="mte:Q" msd="UPosTag=PART|Polarity=Neg" lemma="ne" xml:id="P1.crit.s1.t5">ne</w>
   <w ana="mte:Vmer3s" msd="UPosTag=VERB|Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin"
    lemma="obuti" join="right" xml:id="P1.crit.s1.t6">obuje</w>
   <pc ana="mte:Z" msd="UPosTag=PUNCT" xml:id="P1.crit.s1.t7">.</pc>
   <linkGrp type="UD-SYN" targFunc="head argument" corresp="#P1.crit.s1">
     <link ana="ud-syn:case" target="#P1.crit.s1.t2 #P1.crit.s1.t1"/>
     <link ana="ud-syn:obl" target="#P1.crit.s1.t6 #P1.crit.s1.t2"/>
     <link ana="ud-syn:expl" target="#P1.crit.s1.t6 #P1.crit.s1.t3"/>
     <link ana="ud-syn:nsubj" target="#P1.crit.s1.t6 #P1.crit.s1.t4"/>
     <link ana="ud-syn:advmod" target="#P1.crit.s1.t6 #P1.crit.s1.t5"/>
     <link ana="ud-syn:root" target="#P1.crit.s1 #P1.crit.s1.t6"/>
     <link ana="ud-syn:punct" target="#P1.crit.s1.t6 #P1.crit.s1.t7"/>
   </linkGrp>
 </s>
</seg>
```

AVTOMATSKO JEZIKOSLOVNO OZNAČEVANJE

Za posodabljanje besed v zbirki smo uporabili orodje cSMTiser; orodje približa zapis besed sodobni slovenščini, vendar ne brez napak.
Na osnovi avtomatsko posodobljenih besed je bil korpus nato tudi jezikoslovno označen: lematizacija, oblikoskladenjsko označevanje, skladenjska razčlenitev povedi po formalizmu Universal Dependencies.

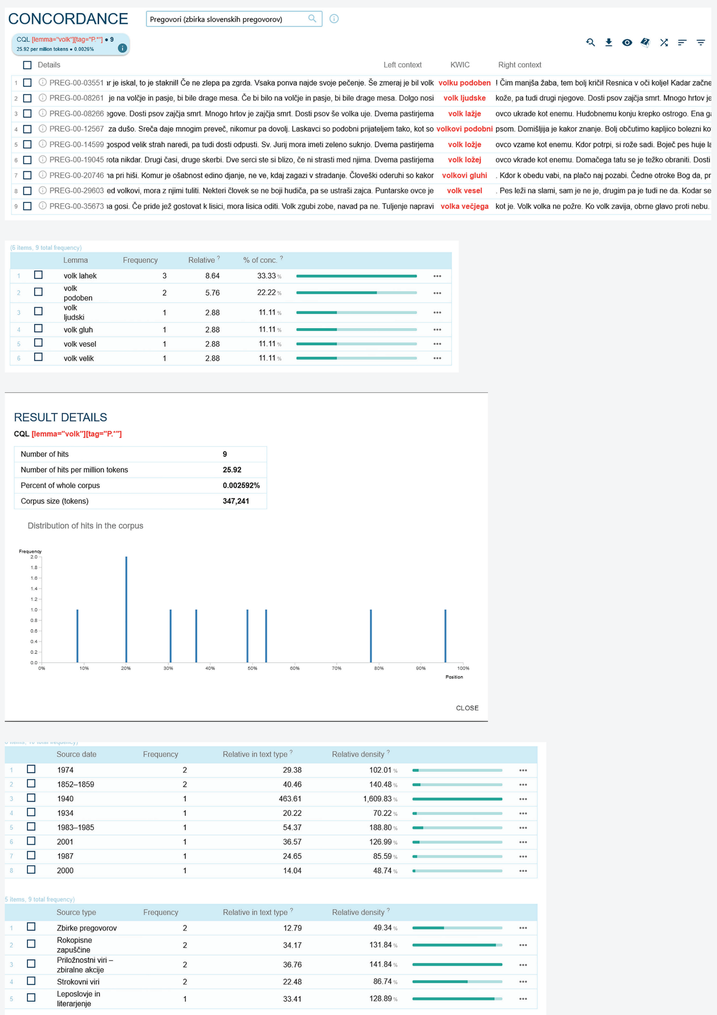
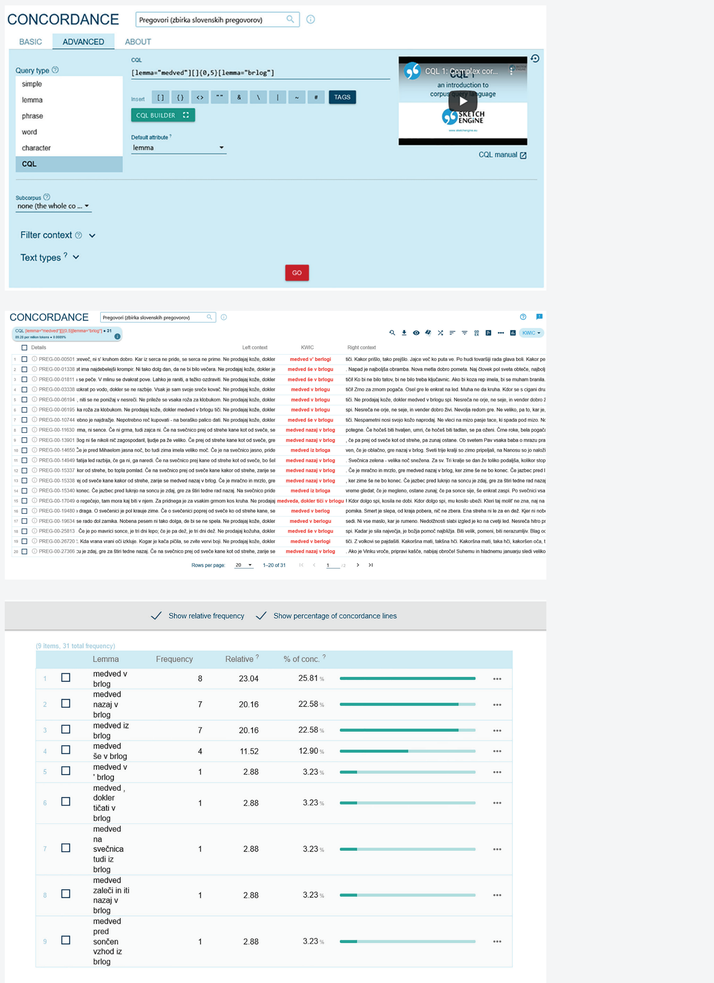
Primer jezikoslovnih oznak v besedilu, npr. za besedno obliko »čevlja«

- Kontekstno odvisna oblikoskladenjska oznaka po priporočilih MULTEXT-East (Erjavec, 2012), tu »Ncmsg« za »Noun Type=common Gender=masculine Number=singular Case=genitive« (obstaja tudi ekvivalentna oznaka v slovenščini, tu »Somer« in njena razširitev v pare lastnost=vrednost);
- Lema oz. osnovna oblika besede, tu »čevelj«;
- Oblikoskladenjske oznake po sistemu Universal Dependencies za slovenski jezik (Dobrovoljc et al., 2017), tu »NOUN Case=Gen Gender=Masc Number=Sing«. Te oznake so podobne oznakam MULTEXT-East, vendar z drugače izpisanim naborom lastnosti in vrednosti, občasno se pa od njih tudi sistemsko razlikujejo;
- Odvisnostna skladenjska razčlenitev povedi po sistemu Universal Dependencies.

ANALIZA ZBIRKE

- **Največje število paremioloških enot je iz zbirke Pregovori in reki na Slovenskem, ki jo je uredil Etbin Bojc (4.884 enot).**
- **Najstarejši pregovori v zbirki so iz leta 1587, in sicer iz Predgovora k Postili Jurija Juričiča.**
- **Samostalnik *dan* : 1.657 pojavnosti**
- **Glagol *biti* : 19.301 pojavnosti**
- **Pridevnik *dobro* : 1.367 pojavnosti**
- **Predlog *v* : 4.538 pojavnosti**

CQL [lemma="volk"][tag="P.*"]



CQL [lemma="medved"][]{0,5}[lemma="brlog"]