Izdelava in analiza digitalizirane zbirke paremioloških enot

Saša Babič*, Tomaž Erjavec†
* Inštitut za slovensko narodopisje ZRC SAZU
Novi trg 2, 1000 Ljubljana
sasa.babic@zrc-sazu.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Članek obravnava digitaliziranje zbirke slovenskih pregovorov Inštituta za slovensko narodopisje ZRC SAZU. Zbirka je nastajala od leta 1947 dalje, digitalizacija pa se je začela v samem začetku 21. stoletja z iniciativo Marije Stanonik. V predstavljenem delu smo izhajali iz Excel razpredelnic paremioloških enot in virov, iz katerih smo najprej izločili neustrezne enote in neuporabljene vire. Nato smo tabeli pretvorili v zapis TEI in pregovore avtomatsko jezikoslovno označili. Tu so bile besede posodobljene, lematizirane, oblikoskladenjsko označene, povedi pa skladenjsko razčlenjene po formalizmu Universal Dependencies. Kanonični zapis TEI smo pretvorili v več izvedenih formatov in zbirko objavili pod odprto licenco na repozitoriju CLARIN.SI, kjer jo je mogoče prevzeti, in na konkordančnikih CLARIN.SI, ki so primerni za jezikoslovne analize zbirke. V članku orišemo tudi način iskanja po zbirki v konkordančnikih, ki omogočajo temeljitejšo etnolingvistično in semiotično raziskavo.

Creation and analysis of a digitised collection of Slovenian paremiological units

The article discusses the digitization of the collection of Slovenian proverbs from the Institute of Slovenian Ethnography ZRC SAZU. The collection was created from 1947, and its digitization began at the start of the 21st century on the initiative of Marija Stanonik. The departure point of the presented were two Excel spreadsheets with paremiological units and their bibliographical sources, from which we removed inappropriate units, and unused sources. The two spreadsheets were then converted to a TEI encoding, and the paremiological units automatically linguistically annotated: words were modernised, lemmatised, morphosyntactically annotated and the sentences syntactically parsed according to the Universal Dependencies formalism. We converted the canonical TEI encoding into several derived formats and published the collection under an open licence on the CLARIN.SI repository, where it can be downloaded, and on the CLARIN.SI concordancers, which allow for linguistic analyses of the collection. The paper also outlines searching the collection in the concordancers, which enables detailed ethnolinguistic and semiotic research.

1. Uvod

Jezik je ohranjevalec in nosilec kulture, s katerim človeštvo ustvarja in vključuje refleksije o samem sebi (Pitkin, 1972; Bartmiński, 2005; Tolstaja, 2015). Ena od najpogosteje rabljenih jezikovnih oblik so pregovori oz. paremiološke enote.

Paremiološke enote ali pregovori v širšem pomenu so eden od najkrajših žanrov slovstvene folklore; pregovore lahko opišemo kot relativno stalne povedi, ki jih uvrščamo med kratke folklorne obrazce. Pogosto so označeni z besednimi zvezami, kot »modrost ljudstva« (Mieder, 1993), »stara modrost« in »poezija vsakdanjega jezika« (Matičetov, 1956). V vsakem primeru lahko trdimo, da so pregovori »skrčeni moralno-etični obrazci določene skupnosti; so neke vrste tradicionalni stereotipi njenega samozavedanja in samoidentifikacije, bili so iz generacije v generacijo prenašani jezik vsakdanje kulture« (Kržišnik, 2008: 38). Prav zato velja, da so pregovori kratki stereotipi na sentenčni ravni s prenesenim ali generalizirajočim pomenom ter so načeloma splošno znani (Grzybek, 2012).

Pregovori so kulturna besedila z velikim semantičnim potencialom (Grzybek, 2015), saj gre za »zaključene misli« (Mlacek, 1983: 131), vendar pa se ne razlikujejo le po besedilu, temveč tudi glede na teksturo in kontekst (Dundes, 1965). Zaradi prozodičnih značilnosti si jih je lažje zapomniti, dandanes pa zato ponujajo možnosti za nadaljnjo uporabo, na primer pri oglaševanju, sodobnem prenosu mnenj, grafitih ali modifikacijah v različnih medijih. Semiotična kompleksnost pregovorov in prepletenost med sintaktično (kratkost), pragmatično (prenašanje skozi različne generacije) in semantično (stereotipno, splošno znanje) razsežnostjo ponujajo raziskovanje pregovorov kot kulturni znak, ki ohranja zgodovino kulture oz. družbe, hkrati pa sprejema nove funkcije, ki širijo in porajajo nove kontekste. Prav zato so paremiološke enote oz. pregovori označeni za narodni zaklad, neprecenljivo modrost in dediščino prednikov, in ne preseneča, da so (bili) predmet sprotnega terenskega zapisovanja ali celo namenskega zbiranja (Arewa in Dundes, 1966; Stanonik, 2015) ter analiz rabe (Meterc, 2021).

Inštitut za slovensko narodopisje ZRC SAZU je sistematično gradil arhiv različnih žanrov slovstvene folklore, v sklopu katerega je nastajala tudi zbirka pregovorov. Ti so bili zabeleženi na kartotečnih listkih ali v tematskih arhivskih mapah. V začetku 21. stoletja se je pojavila potreba po digitalizaciji gradiva, ki bi omogočala lažje delo z gradivom.

Pri projektu *Tradicionalne paremiološke enote v dialogu s sodobno rabo* (2020–2023) smo predvideli združitev etnolingvističnih pristopov in semiotike z namenom diahronega vpogleda v družbo s pomočjo pregovorov. Da bi bila analiza temeljitejša, je pomemben del projekta pretvorba gradiva v sprejemljivo obliko za računalniško besedilno analizo.

V članku opišemo pripravo in jezikoslovno označevanje digitalizirane zbirke pregovorov, ki je sedaj dostopna na repozitoriju in konkordančnikih CLARIN.SI, ter uporabo digitalizirane zbirke v namene etnolingvistične obravnave paremioloških enot. Na koncu podamo zaključke in načrte za nadaljnje delo.

2. Priprava gradiva

Inštitut za slovensko narodopisje (ISN) ZRC SAZU v arhivu hrani folklorno gradivo v analogni obliki, tj. ročno napisani, natipkani ali natisnjeni na kartotečnih listkih, v arhivskih predalih in omarah. Težnja po digitalizaciji folklornega gradiva se je najprej začela pri pregovorih, za katere je Marija Stanonik že leta 1997–1999 pridobila projekt *Slovenski pregovori in rekla* (Stanonik, 1996), v katerem je začela širiti arhivsko zbirko pregovorov na ISN. Z mislijo na digitalizacijo je nadaljevala v projektih *Informatizacija neoprijemljive dediščine za etnologijo in folkloristiko* (2005–2008) (Stanonik, 2004) in *Slovenski pregovori kot kulturna dediščina: klasifikacija in redakcija korpusa* (2010–2013) (Stanonik, 2009; Stanonik, 2015). Gradivo je bilo dodano k obstoječi zbirki v računalniškem prepisu, sprva v programu Word, pozneje v programu Excel, kar je predstavljalo temelj, na katerem smo lahko izvedli pretvorbo v druge digitalne formate.

2.1. Priprava gradiva v razpredelnicah

V urejanje smo dobili dve excelovi tabeli: prva je vsebovala 59.543 večinoma paremioloških enot, druga pa 2.742 virov teh enot. Tabeli sta bili povezani s kodo, ki je bila določena viru. Ob pregledu gradiva smo ugotovili, da precej enot ne spada v paremiološki nabor; te smo ročno izločili (uganke, dele folklornih pesmi, pozdrave, frazeme ipd.), pri pregledu virov smo ročno izločili vse tiste, ki niso bili navedeni ob paremioloških enotah. Poleg tega so nekatere paremiološke enote vsebovale širši kontekst, ki smo ga ročno izbrisali; tako smo dobili poenoteno obliko samostojnih paremioloških enot. Pri vremenskih paremioloških enotah se je pojavil problem pojasnjevanja svetniškega poimenovanja dnevov in praznikov: v originalnem zapisu (časopisi, koledarji, zvezki ipd.) so bili navedeni kot pojasnilo, npr. *Če je na Velike maše dan [15. avgust] lepo vreme, potem bo ozimna pšenica lepa*; *Če na ta dan [Florijanovo, 4. maj] dež gre, potlej ga celo leto manjka*. V excelovi tabeli, ki predstavlja del Inštitutskega arhiva, smo te pustili zabeležene v oglatem oklepaju.

Po ročnem urejanju smo Excel dokumente združili z OpenRefine[1] in tako poenotili korektorske opombe in kategorije označevanja pregovorov. Osnovne popravke smo vnesli tudi pri preverjanju shematiziranih vnosov (npr. navajanje virov, odstranjevanje presledkov na koncu besedil v posameznih celicah ipd.). Sledil je prenos podatkov v delovno bazo SQLite[2], kjer so potekali popravki preprostih slovničnih napak in zatipkov (velike začetnice, dvojni presledki, nepravilna raba ločil ipd.) ter zaznava uporabljenih črkovnih naborov, kjer gre izpostaviti nestandardizirane zapise dajnčice, metelčice, bohoričice in gajice. Pregovore so namreč začeli prepisovati v računalniško obliko že v začetku 21. stoletja, ko nabor črkovnih znakov še ni bil tako pester in so prepisovalci reševali zagate z različnimi zapisi z improviziranim izborom znakov. Po osnovnih popravkih paremioloških enot smo nadaljevali z iskanjem enakih oz. podvojenih enot in odstranjevali dvojnike, pri čemer smo vse vire dodali k eni paremiološki enoti. Ob koncu urejanja smo podatke izvozili v format TSV (tab-separated values), ki je bil izhodišče za izdelavo korpusa.

Gradivo je tako po ročnem in strojnem urejanju vsebovalo 36.349 relativno enotno urejenih paremioloških enot ter 2.515 virov.

Razpredelnica z viri vsebuje za vsak bibliografskimi 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 enote (torej zapis enote, kot se pojavi v viru) in kritično transkripcijo enote, ki enote, zapisane v bohoričici, transkribira v gajico. Tako ima npr. enota PREG-00-00001 zaporedno številko 1, seznam virov bib14.1: 202; bib23.1: 51; bib7.1: 524, diplomatično transkripcijo »Bres muje ſe zhreul ne obuje.« in kritično transkripcijo »Brez muje se črеul ne obuje.«

2.2. Zapis TEI

V naslednjem koraku smo podatke iz dokumentov TSV pretvorili v zapis, ki je bolj primeren za hrambo kot tudi za nadaljnje obdelave, in sicer XML s shemo po priporočilih iniciative za kodiranje besedil TEI (TEI Consortium, 2020). Celotna zbirka je bila formirana kot en TEI dokument (element <TEI>) s kolofonom (element <teiHeader>) in besedilnim delom (<text>).

Kolofon vsebuje bibliografske in druge metapodatke o zbirki, kot je npr. taksonomija kategorizacije virov. V opisu vira (<sourceDesc>) vsebuje tudi celoten seznam virov paremioloških enot; zapis je ilustriran v sliki 1.

Besedilni del vsebuje paremiološke enote, vsako s svojim identifikatorjem, diplomatični in kritični prepis ter seznam njihovih virov; zapis ilustriramo v sliki 2.

2.3. Posodabljanje besed in drugo jezikoslovno označevanje

Precejšnjo težavo za uporabo izdelane zbirke je predstavljal zapis v arhaični slovenščini, ki oteži iskanje po pregovorih, kot tudi njihovo nadaljnjo analizo. Oteženo je tudi avtomatsko jezikoslovno označevanje zbirke, saj orodja za jezikoslovno označevanje delujejo dobro le na sodobni standardni slovenščini.

Za posodabljanje zbirke smo uporabili odprtokodno[3] orodje za normalizacijo cSMTiser (Scherrer in Ljubešić, 2016), ki temelji na principu statističnega strojnega prevajanja in orodju Moses (Koehn, 2010). cSMTiser smo naučili posodabljanja na ročno posodobljene korpusu slovenščine goo300k (Erjavec, 2016), podobno, kot smo že pred tem naredili za posodabljanje zbirke slovenskih romanov v okviru korpusa ELTeC (Schöch et al., 2021). Z orodjem smo nato normalizirali kritični prepis, pri čemer orodje sicer približa zapis besed sodobni slovenščini, dela pa tudi napake (npr. besedo »črеul« prevede v »čevlj« namesto »čevelj«).

[1] https://openrefine.org/
[2] https://www.sqlite.org/
[3] https://github.com/clarinsi/csmtiser

```
<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>
```

Slika 1: Primer virov paremioloških enot v zapisu TEI.

```
<listBibl xml:lang="sl">
    <bibl xml:id="bib1.1" n="15" corresp="#bibl.dictionary">Bohorič,
        Adam, <date type="reprint" when="1970">1970</date>
        (<date type="firstEdition" when="1584">1584</date>):
        Arcticae horulae succisivae. Faksimile. Mladinska knjiga.
        Ljubljana.</bibl>
    <bibl xml:id="bib2.1" n="3" corresp="#bibl.dictionary">Kastelec, Matija,
        (<date from="1680" to="1685">1680-1685</date>): Dictionarivm
        Latino-Carniolicvm. NUK, rokopisna zbirka, MS 803. Ljubljana.</bibl>
    <bibl xml:id="bib54.2" n="901/9" corresp="#bibl.yearbook">Erjavec, Fran,
        (<date when="1880">1880</date>): Iz potne torbe. V: Letopis Matice
        Slovenske za leto 1880. Matica Slovenska. Ljubljana.</bibl>
```

Slika 2: Primer kodiranja paremiološke enote v zapisu TEI.

```
<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>
```

Slika 3: Primer kodiranja jezikoslovno označene paremiološke enote v zapisu TEI.

Na osnovi avtomatsko posodobljenih besed smo nato korpus jezikoslovno označili. Tu smo uporabili odprtokodno orodje CLASSLA[4] (Ljubešić in Dobrovoljc, 2019), s katerim smo dodali naslednje jezikoslovne oznake v besedilo, npr. za »čevlja«:

- oblikoskladenjsko oznako po priporočilih MULTEXT-East (Erjavec, 2012), npr. »Ncmsg« za »Noun Type=common Gender=masculine Number=singular Case=genitive« (pri čemer obstaja tudi ekvivalentna oznaka v slovenščini, tu »Somer« in njena razširitev v pare lastnost=vrednost);
- lemo oz. osnovno obliko besede, tu »čevelj«;
- oblikoskladenjske oznake po sistemu Universal Dependencies za slovenski jezik (Dobrovoljc et al., 2017), npr. »NOUN Case=Gen Gender=Masc Number=Sing«. Te oznake so sicer podobne oznakam MULTEXT-East, vendar z drugače izpisanim naborom lastnosti in vrednosti, občasno se pa od njih tudi sistemsko razlikujejo;
- odvisnostno skladenjsko razčlenitvijo povedi po sistemu Universal Dependencies.

Jezikoslovno označena različica posamezne paremiološke enote je bila dodana v zapis TEI po njenih besedilnih zapisih; format je ilustriran v sliki 3. V različici korpusa, ki vsebuje posodobljene in jezikovno označene enote, je dopolnjen tudi kolofon s taksonomijo skladenjskih oznak Universal Dependencies in z opisom uporabljenih orodij.

[4] https://github.com/clarinsi/classla

2.4. Objava zbirke

Zbirko smo objavili na dva načina. Za prevzem je dostopna na repozitoriju CLARIN.SI (Babič et al., 2022) pod odprto licenco CC BY. Poleg obeh različic zbirke (brez in z jezikoslovno označenimi enotami) v formatu TEI je tam na voljo tudi v izvedenem formatu TSV, torej kot razpredelnici z viri in enotami, in v t. i. vertikalnem formatu, ki služi kot vhodni format za konkordančnike CLARIN.SI.

Zbirka je dostopna tudi na konkordančnikih noSketch Engine in KonText CLARIN.SI. Prek teh dveh storitev je omogočen analitični vpogled v digitalizirano zbirko.

3. Analiza gradiva

Zbirka ima v diplomatičnem zapisu zavedenih 36.066 paremioloških enot (283 jih je v kritičnem prepisu). Največ paremioloških enot je izpisanih iz že obstoječih zbirk pregovorov (10.187 enot) ter iz leta 1974, tj. zbirke pregovorov *Pregovori in reki na Slovenskem*, ki jo je uredil Etbin Bojc (4.884 enot). Treba je upoštevati dejstvo, da je Bojc precej paremioloških enot zbral tudi iz že prej obstoječih zbirk (npr. Kocbek (1887), Kocbek-Šašelj (1934) ter starejše slovnice in slovarji), velja pa njegova zbirka za prvo sodobnejšo. Najstarejši pregovori v zbirki so iz leta 1587, in sicer iz *Predgovora k Postili* Jurija Juričiča.

Zbirka paremioloških enot ISN vsebuje precej enot iz slovnic in slovarjev, kar pomeni, da so bile te enote zapisane kot izolirane entitete, brez konteksta. Poleg tega, navedeno ne izpriča dejanskega poznavanja in rabe paremioloških enot, kot ga lahko predvidevamo pri zbiranju paremiološkega gradiva na terenu ali iz tiskanih besedil, v katerih avtor predvideva poznavanje posameznih paremioloških enot in s tem bralčevo razumevanje napisanega. Navedeno je v folkloristiki pomemben del raziskav in analiz, saj razkriva tudi konceptualni in etnolingvistični vidik folklornega gradiva. Če predvidimo dobro poznavanje posameznega pregovora (npr. *Brez muje se še čevelj ne obuje*), lahko predvidimo tudi konceptualno ozadje in etnolingvistično sliko, ki nam jo tovrstno gradivo lahko ponudi. Za takšen vpogled se poslužujemo ne le etnolingvističnega pristopa (povezovanja jezikoslovja in etnologije s poudarkom na stereotipni predstavi pojava), temveč tudi semiotične analize (pomen znaka).

3.1. Etnolingvistična in semiotična analiza s pomočjo konkordančnikov

Čeprav pregovori tradicionalno spadajo na področje paremiologije, so pogosto tudi raziskovalni predmet folkloristike, sociologije, pedagogike, jezikoslovja itd. Semiotika, kot veda o znakih, ponuja metodologijo za raziskovanje globljih dimenzij prepletenih kulturnih ozadij pregovorov (Grzybek, 2014). Semiotika s poudarkom na pragmatični (razmerje med označenim in označencem), sintaktični (formalni odnosi med znaki) in semantični dimenziji (odnosi znakov s predmeti, za katere je mogoče uporabiti znake) (Morris, 1938) omogoča opazovanje pregovorov z globljim vpogledom v kulturne pomene, pojme in svetovne nazore. Do svetovnega nazora v pregovorih pa je moč dostopati z etnolingvističnimi raziskovalnimi metodami, vključno z diahronim in sinhronim pristopom.

Etnolingvistika kot samostojno področje daje jeziku posebno mesto v družbi: v jeziku se oblikujejo kulturni pomeni; jezik v besedah, frazeologiji, celo v slovnici posreduje podobe sveta. Jezik je s tega vidika »gradivo kulture«, medtem ko je hkrati tudi kulturni meta-jezik: skupaj s folkloro velja za enega ključnih kulturnih kodov in kulturno ekspresivnih oblik. Jezik je zato eden najpomembnejših virov za raziskovanje folklore in rekonstrukcij njenih zgodnjih stanj; povezava med jezikom in kulturo je vzajemna (Tolstaja, 2006) in skupaj tvorita znakovni sistem. Vsi kulturni pomeni se zberejo v semantiko poimenovanja z besedami; te je lublinska etnolingvistična šola poimenovala jezikovni stereotipi (Bartmiński, 2005), ki kažejo naš poskus nadzora sveta. Analize relativno stalnih besednih zvez in besed v določenih kontekstih nam prikazujejo jezikovni zemljevid sveta z najpomembnejšimi družbenimi podobami in predstavami.

Hitro razvijajoče se področje digitalne humanistike omogoča raziskovalcem sprejemanje novih, korenito drugačnih metod raziskovanja in, kar je prav tako pomembno, daje na voljo elektronske zbirke z naprednimi možnostmi iskanja podatkov (Rassmusen Neal, 2015). Korpusno jezikoslovje in trenutno priljubljena »metodologija branja na daljavo« (tj. uporaba e-virov) poskuša izkoristiti velike jezikovne vzorce, da bi pridobili (kvantitativni) vpogled v besedišče, uporabo, trende in vizualizacije na področjih jezikovnega interesa. Hkrati pa takšne računalniške tekstovne oblike zbirk omogočajo natančnejše in hitrejše kvalitativne analize večjih zbirk: posameznih konkordančnih kombinacij in besednih okolij.

Semiotična analiza v namene etnolingvistične raziskave (Bartmiński, 2005) paremiološkega gradiva poteka predvsem na ravni semantike: pri besedah želimo zaznati tako metaforične pomene kot stereotipne oznake, ki jih (posamezna) beseda vsebuje in hkrati posreduje prek metafore v širši kontekst, torej s semiotičnega vidika, kakšni znaki se tvorijo znotraj paremiološke enote.

Statistični vpogled v celotno zbirko pokaže med drugim tudi najbolj pogosto rabljene besede, ki lahko podajo tudi splošnejša predvidevanja o družbeni naravnanosti. Najpogostejša polnopomenska beseda v zbirki paremioloških enot je:

- samostalnik *dan* se pojavi 1.657-krat; ta metaforično ali metonimično označuje tako časovno omejeno obdobje, ki pomeni dolgo (*Premislek je boljši kot dan hoda.*) ali kratko (*Bitke ne dobiš v enem dnevu.*), konec obdobja (*Po večeru se dan pozna.*), poimenovanje konkretnega dneva (*Ni vsak dan praznik./Pavla dne lepo, leto dobro bo.*), sledenje dobrega oz. označevanje konceptualne cikličnosti (*Za vsako nočjo pride dan*). Najpogostejša pojavnost ne presenеča, saj je ta samostalnik zelo pogost sestavni element vremenskih in kmetijskih paremioloških enot, poleg tega pa je je tudi v splošnem sodobnem jeziku izredno pogost: v Gigafidi v2.0 je tretji najpogosteje rabljeni samostalnik[5]. Po drugi strani je smiselno izpostaviti, da se nasprotje, tj. *noč* pojavi le 318-krat (pojavlja se kot nasprotje dnevu (*Ljubezen vidi noč, kjer sije beli dan.*), temen čas, ko se ne vidi (*Ponoči so vse krave črne.*), vpliven čas (*Noč ima svojo moč.*), mejni čas (*Ne hvali*

[5] http://hdl.handle.net/11346/QHKH

dneva pred nočjo.), slab čas (*Dan se zjutraj išče, noč pa sama pride.*), oznaka prazničnih časov *velika noč, božična noč*) itd.).

- Glagol *biti* se pojavi 19.301-krat (zanikan pa 3.501-krat), kar ne preseneča, glede na to, da gre za enega najosnovnejših glagolov; glagol je najpogostejši tudi v splošnem sodobnem jeziku.[6]
- Pridevnik *dobro* se pojavi 1.367-krat, največkrat v osnovni obliki, najmanjkrat pa kot presežnik (prim. *slabo* se pojavi 301-krat, osnovnik najpogosteje, presežnik najmanjkrat). Na podlagi izoliranih enot bi lahko sklepali, da pregovori na semantični ravni pogosto izražajo vrednotenje stanja ali delovanja, kar poleg izražanja družbenega nazora potrjuje tudi njihov pedagoški potencial.
- Predlog *v* je najpogostejši predlog v paremioloških enotah, tj. pojavi se v 4.538-krat. Iz tega podatka lahko sklepamo, da izvorno konceptualno najpogosteje uvrščamo pojave znotraj časovno-prostorskega koncepta pojava, pa čeprav se pomensko raba predloga razširi tudi na izražanje namena, sredstva, odnosa do celote, dejanja/stanja ipd. Enako je opaziti tudi v sodobnem splošnem jeziku.[7]

Ob najpogostejši prisotnosti besed v paremioloških enotah *dan*, *biti*, *dobro* in *v* se izkaže, da te povsem ustrezajo tudi pogostnosti rabe v splošnem sodobnem jeziku, ne glede na to, da gre za večinoma arhivsko gradivo.

Za natančnejši etnolingvistični in konceptualni vpogled je primernejša analiza s posamezno sestavino (npr. samostalnikom *čevelj*, *medved*) in njenimi vezavami, na podlagi katerih lahko s semiotično metodo podamo interpretacije družbenih konceptualnih vidikov. Za tako analizo je najširše uporabno enostavno iskanje, ki v primeru te zbirke naniza vse sklone iskanega samostalnika, vključno s starejšimi zapisi, npr. pri iskanju besede *čevelj* (68 enot) iskalnik izloči vse sklone, prav tako pa zapis *črevelj*, *čevl* ipd. Ob zahtevnejših iskanjih je možno slediti tudi številu posamezni obliki zapisa: *črevelj* (2), *črevlju* (1), *čevle* (3), seznam besed pa omogoča tudi sledenju starejšim zapisom, virom in njihovi pogostnosti v časovnem razponu, variantnim rabam in morebitnim prenovitvam.

Enako pri iskanju vseh zapisov in sklonov besede *lisica* (starejša oblika *lesica*, 7 enot) iskalnik najde 93 paremioloških enot. Ob zahtevnejših iskanjih je možno slediti tudi številu posamezni obliki zapisa: *lisica* (31), *lisice* (5), *lisici* (7), *lisico* (6), *lesica* (7), itd. Kontekstualna raziskava objav po različnih virih poda poveden podatek: slovnice in slovarji navajajo paremiološke enote z besedo *lisica*, ki so v celoti metaforične in se nanašajo na ljudi, medtem ko koledarji navedejo tudi paremiološke enote, ki veljajo za vremenske napovedi.

Iskalnik omogoča tudi iskanje želene besede v navezavi z drugo besedno vrsto, npr. lema *medved*, ki mu sledi glagol. Sicer je tako moč ugotoviti marsikatere povezave, vendar sam statistični del v nasprotju s pričakovanji prikaže tudi rezultate iz drugih (predhodnih ali sledečih) pregovorov, ne le rezultate, vezane na posamezni pregovor. Na primeru besede *medved* statistični del prikaže 79 ustreznic, vendar je teh znotraj enega pregovora 66. Ob ročnem pregledu kaj hitro ugotovimo, da se ta beseda najpogosteje veže z glagolom *prodajati*. Ob navezavah na samostalnik se pojavlja *koža*, kar tvori pregovor, ki metaforično svari pred preuranjeno hvalo. Pregovor nakazuje semantično polje, ki se v etnolingvistični interpretaciji veže na ekonomski odsev družbe, tj. prodaje medvedove kože, ki v zgodovinskem kontekstu pokaže svojo veliko ekonomsko vrednost.

Kljub vsemu iskalnik zaradi starejših in narečnih izrazov ne poišče vedno vseh kombinacij, npr. pri *Lep čevelj vidiš, a ne veš, kje me gloje* ali *Kdor stare čevlje flika, pride do zlatnika*, kjer konkordančnik ni zaznal kombinacije samostalnika in glagola.

Variante posameznega pregovora najlažje najdemo z iskanjem po besednih zvezah, npr. iskanje besedne zveze *lastovka ne* poda štiri rezultate: *Ena lastovka ne naredi poletja*, *Ena lastovka ne naredi pomladi*, *Ena lastovka ne prinese pomladi*, *Ena lastovka ne prinese nikoli spomladi*. Pri glagolski besedni zvezi *gre samo enkrat na led* pa rezultat poda tako *osla* kot *lisico* (*Osel/lisica gre samo enkrat na led*), prav tako *svoj rep hvali* lahko tako *lisica* kot *mačka* (*Vsaka lisica/mačka svoj rep hvali*).

Etnolingvistični vpogled v korpus pregovorov je z digitaliziranim gradivom in možnostjo zahtevnejšega iskanja temeljitejši. Že pogostost posameznih besed v pregovorih ali pa podatek o variantnosti posameznega pregovora je odlično izhodišče, ki ga z analognim arhivom le težko dosežemo.

4. Sklep

Digitalizacija folklornega gradiva olajša analizo le tega, hkrati pa postane bistveno bolj natančna – iskalniki omogočajo izpis vseh želenih enot, hkrati pa je primerjava gradiva bolj dosledna.

Vzpostavitev digitalne zbirke paremioloških enot ISN pomeni premik v slovenski folkloristiki. Gradivo je dostopnejše in analitično lažje obvladljivo. Hkrati takšna oblika ne terja (semantične, tematske, funkcijske, abecedne ipd.) kategorizacije pregovorov, temveč so razvrščeni kot najmanjša zaključena besedila, na katerih izvedemo analizo. Nedvomno je glede problema kategorizacije takšna rešitev najugodnejša, saj sama kategorizacija pogosto pokaže več pomanjkljivosti kot prednosti.

Pri zbirki pregovorov vsekakor najdemo mesto za izboljšave: poleg odprave nekaterih pravopisnih napak, se poraja vprašanje variantnosti ter povezave med variantami; na ta način bi bili odstranjeni tudi še nekateri podvojeni pregovori (predvsem tisti, ki so vpisani z različnimi ločili, npr. eden z vejico, drug s podpičjem). Ker je ponekod diplomatični prepis problematičen (gajica, bohoričica, metelčica, fonetični zapis), se poraja vprašanje smiselnosti knjižnega zapisa pregovora, ki bi moral biti ročno preverjen. Zbirka bo vsekakor tudi dopolnjena z novimi paremiološkimi enotami (iz starejših virov kot sodobne rabe). Poleg teh pa bi bilo smiselno uvesti tudi razdelitev virov po kategorijah, ki bi natančneje prikazal prisotnost paremioloških enot v posamezni kategoriji virov, kar bi omogočalo tudi primerjalno analizo (npr. enote v koledarjih in enote v slovnicah).

Za izdelavo digitalne paremiološke zbirke smo posegli po sistemih, ki so ustaljeni v jezikoslovju. V premisleku pa ostaja, kako digitalizirati slovstveno folkloro, ki je daljša (npr. zgodbe, molitve) in ima specifične funkcije (npr. uganke, zagovori).

[6] http://hdl.handle.net/11346/XNRI

[7] http://hdl.handle.net/11346/ZYVZ

Zahvala

Digitalizirana zbirka paremiolških enot ne bi nastala brez projektnih sodelavcev, še posebej Miha Pečeta: njegov občutek za folklorno gradivo in poznavanje računalniškega sveta sta omogočila hiter potek dela in sprotno reševanje zagat.

Delo opisano v prispevku je podprl temeljni raziskovalni projekt »Tradicionalne paremiološke enote v dialogu s sodobno rabo« (ARRS J6-2579).

5. Literatura

Ojo Arewa in Alan Dundes. 1966. Proverbs and the Ethnography of Speaking Folklore. *American Anthropologist*, 64: 70–85.

Saša Babič, Miha Peče, Tomaž Erjavec, Barbara Ivančič Kutin, Katarina Šrimpf Vendramin, Monika Kropej Telban, Nataša Jakop, in Marija Stanonik. 2022. *Collection of Slovenian paremiological units Pregovori 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1455.

Jiři Bartmiński. 2005. *Jazykovoj obraz mira: očerki po etnolingvistike*. Indarik, Moskva.

Kaja Dobrovoljc et al. 2017. The Universal Dependencies Treebank for Slovenian. V: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, str. 33–38, Association for Computational Linguistics, doi:10.18653/v1/W17-1406.

Alan Dundes. 1965. *The study of folklore*. Prentice-Hall, Englewood Cliffs.

Tomaž Erjavec. 2021. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1): 35–57.

Tomaž Erjavec. 2015. *Reference corpus of historical Slovene goo300k 1.2*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1025.

Diana Faridovna Khakimzyanova in Enzhe Kharisovna Shamsutdinova. 2016. Corpus Linguistics in Proverbs and Sayings Study: Evidence from Different Languages. *The Social Sciences*, 11(15): 3770–3773.

Peter Grzybek. 2012. Proverb Variants and Variations: A New Old Problem? V: O. Lauhakangas, ur., in R. J. B. Soares, ur., *Proceedings of the Fifth Interdisciplinary Colloquium on Proverbs*, str. 136–152, AIP-IAP, Tavira.

Peter Grzybek. 2014. Semiotic and Semantic Aspects of the Proverb. V: H. Hrisztova-Gotthardt, (ur.) in M. A. Varga, ur., *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies*, str. 68–111, De Gruyter, Warsaw/Berlin.

Dell Hymes, D. 1962. The ethnography of speaking. V: T. Gladwin, ur., in W. C. Sturtevant, ur., *Anthropology and Human Behavior*, str. 13–53, Anthropological Society of Washington, Washington.

Fran Kocbek. 1887. *Pregovori, prilike in reki*. Založil Anton Trstenjak, Ljubljana.

Fran Kocbek in Ivan Šašelj. 1934. *Slovenski pregovori, reki in prilike*. Družba Sv. Mohorja, Ljubljana.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Erika Kržišnik. 2008. Kulturološka interpretacija frazema. V: M. Kalin Golob, ur., N. Logar Berginc, ur., in A. Grizold, ur., *Jezikovna prepletanja*, str. 149–165, Fakulteta za družbene vede, Ljubljana.

Nikola Ljubešić in Kaja Dobrovoljc. 2019. What Does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, str. 29–34, Association for Computational Linguistics, doi:10.18653/v1/W19-3704.

Milko Matičetov. 1956. Pregovori in uganke; ljudska proza. Slovenska matica, Ljubljana.

Matej Meterc. 2021. Aktualna raba in pomenska določljivost 200 pregovorov in sorodnih paremioloških izrazov. *Jezikoslovni zapiski* 27(1): 45–61.

Jozef Mlacek. 1983. Problémy komplexného rozboru prísloví a porekadiel. *Slovenská reč* 48(2): 129–140.

Wolfgang Mieder. 1993. *Proverbs are never out of season: Popular wisdom in modern age*. Oxford University Press.

Hanna F. Pitkin. 1972. *The concept of representation*. University of California Press.

Diana Rassmusen Neal. 2015. *Indexing and retrieval of non-text information*. De Gruyter Saur, Chicago, Vancouver.

Yves Scherrer in Nikola Ljubešić. 2016. Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation. V: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, str. 248–55.

Christoph Schöch, Roxana Patraş, Tomaž Erjavec, Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC). *Modern languages open*, doi: 10.3828/mlo.v0i0.364.

Marija Stanonik. 1996. *Slovenski pregovori in rekla*. Projektna prijava.

Marija Stanonik. 2004. *Informatizacija neoprijemljive dediščine za etnologijo in folkloristiko*. Projektna prijava.

Marija Stanonik. 2009. *Slovenski pregovori kot kulturna dediščina: klasifikacija in redakcija korpusa*. Projektna prijava.

Marija Stanonik. 2015. Slovenski pregovori kot kulturna dediščina. Klasifikacija in redakcija korpusa. *Traditiones*, 44(3): 171–214.

Kathrin Steyer. 2017. Corpus Linguistic Exploration of Modern Proverb Use and Proverb Patterns. V: R. Mitkov, ur., *Europhras 2017. Computational and corpus-based phraseology: Recent advances and interdisciplinary approaches. Proceedings of the Conference Volume II*, str. 45–52, London, Geneva.

TEI Consortium. 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. https://tei-c.org/guidelines/P5/

Svetlana M. Tolstaja. 2015. *Obraz mira v tekste i rituale*. Univerza Dimitrija Požarskega, Moskva.