OBLIKOSLOVNI VZORCI V LEKSIKONU SLOLEKS: IZHODIŠČNI NABOR ZA SAMOSTALNIKE

[1]Špela ARHAR HOLDT, [2]Jaka ČIBEJ
[1]Center za jezikovne vire in tehnologije, Univerza v Ljubljani (Filozofska fakulteta, Fakulteta za računalništvo in informatiko)
[2]Inštitut "Jožef Stefan"

Arhar Holdt, Š., Čibej, J. (2018): Oblikoslovni vzorci v leksikonu Sloleks: izhodiščni nabor za samostalnike. Slovenščina 2.0, 6 (2): 33–66.
DOI*:* http://dx.doi.org/10.4312/slo2.0.2018.2.33-66.

Prispevek predstavlja prvi korak k dopolnjevanju leksikona Sloleks z oblikoslovnimi vzorci, in sicer na primeru samostalnikov. Vzorci so v prvem koraku strojno pridobljeni iz leksikona na osnovi izbranih razločevalnih lastnosti (oblikoskladenjskih oznak in spremenljivih delov besednih oblik). Sledi ročno razvrščanje, v katerem (a) ločimo sistemsko in v rabi utemeljene vzorce od primerov, ki se pojavljajo spričo šuma pri strojnem luščenju in nedoslednosti v leksikonu Sloleks; (b) uredimo skupine glede na vsebovanost in sorodnost podatkov; (c) poiščemo in natančneje opredelimo variantnost, tako pri standardnih kot nestandardnih oblikah; (č) začrtamo korake za nadaljnji razvoj programa in nadgradnjo leksikona. Rezultat je izhodiščni nabor formaliziranih oblikoslovnih vzorcev za (občno- in lastnoimenske) samostalnike, ki prinaša 10 skupin za moški spol, 9 skupin za ženski spol in 8 skupin za srednji spol. Priprava nabora vzorcev je razkrila številne možnosti za izboljšavo leksikona, strojno namenski pogled na pregibanje pa priložnosti za dopolnitev slovničnega opisa slovenščine. V nadaljevanju dela bodo vzorci pripravljeni tudi za preostale besedne vrste in dopolnjeni s korpusnim gradivom. Končna nomenklatura bo vpisana v bazo leksikona Sloleks, v obliki strojno berljivih vzorcev pa bo objavljena tudi na repozitoriju Clarin.si.

Ključne besede: Sloleks, leksikon besednih oblik, oblikoslovni vzorci, samostalnik, slovenščina

1 LEKSIKON SLOLEKS

Sloleks je odprtodostopni leksikon besednih oblik za slovenščino,[1] ki poleg osnovne oblike besede vsebuje nabor pregibnih oblik, podatke o pogostosti leme in pregibnih oblik iz referenčnega pisnega korpusa Gigafida (Logar in dr. 2012), zbir standardnih in nestandardnih oblikoslovnih variant ter povezave na besedotvorno sorodne besede. Leksikon je bil pripravljen v projektu Sporazumevanje v slovenskem jeziku[2] po specifikacijah v (Erjavec in dr. 2008), kot je opisano v (Arhar 2009). Trenutno je na voljo popravljena in dopolnjena različica 1.2 (Dobrovoljc in dr. 2015a), ki jo z vidika namena, formata, vsebine in nadaljnjega razvoja natančno opredeljujejo (Dobrovoljc in dr. 2015b).

Sloleks je bil kot vir že večkrat uporabljen pri razvoju jezikovnotehnoloških orodij za obdelavo slovenščine, v prvi vrsti za oblikoskladenjsko označevanje slovenskih besedil (Grčar in dr. 2012, Ljubešić, Erjavec 2016, Čibej in dr. 2016), pa tudi denimo za modernizacijo historičnih slovenskih besed (Scherrer, Erjavec 2013) in normalizacijo slovenskih tvitov (Ljubešić in dr. 2014), za avtomatsko napovedovanje stopnje (ne)standardnosti spletnih besedil (Ljubešić in dr. 2015), za avtomatsko generiranje besednih oblik s pomočjo strojnega učenja (Rejc 2017) ter za luščenje terminologije iz forumskih zapisov (Vintar 2015) in besedil s področja borznega posredništva (Pollak, Božinovski 2014). Leksikon je pogosto rabljen tudi med jezikovnimi uporabniki, ki prek vmesnika iščejo odgovore na jezikovne zadrege, povezane z oblikoslovjem (Dobrovoljc 2015); s tega vidika je bil prepoznan kot dragocen pripomoček za uporabo pri pouku slovenščine (Stritar, Dobrovoljc 2013). V literaturi pa je bilo tudi že opozorjeno, da Sloleks potrebuje nadgradnjo. Ob razvoju koncepta za slovar sodobnega slovenskega jezika (Gorjanc in dr. (ur.) 2015) so bili med informacijskimi dopolnitvami, ki bi tako razvojni kot uporabniški skupnosti najbolj koristile, mdr. izpostavljeni formalizirani oblikoslovni vzorci

[1] Leksikon je dostopen prek vmesnika na spletni strani: http://www.slovenscina.eu/sloleks, kot baza pa v repozitoriju Clarin.si: http://hdl.handle.net/11356/1039 (Dobrovoljc in dr. 2015a).
[2] Projekt je potekal med leti 2008 in 2013, spletna stran: www.slovenscina.eu.

(Dobrovoljc in dr. 2015b: 95):

> Eno najpomembnejših vprašanj, povezanih tako s širitvijo kot reevalvacijo obstoječih oblikoslovnih leksikonov za slovenščino, je izdelava nabora strojno berljivih vzorcev pregibanja besed v slovenskem jeziku, ki bi omogočil validacijo pregibnih paradigem iztočnic v obstoječih priročnikih, pripisovanje paradigem novim lemam ter razvoj metod za njihovo samodejno prepoznavanje v besedilnih korpusih /.../

Odprto dostopni nabor vzorcev, pripravljen namensko za strojno obdelavo slovenščine, bi torej omogočil nadgradnje leksikona in drugih jezikovnih virov ter večjo natančnost označevanja oz. pridobivanja jezikovnih podatkov iz besedilnih korpusov. Na drugi strani bi vključitev vzorcev in lem, ki se po določenem vzorcu pregibajo, v leksikonski vmesnik uporabnikom ponudila pregled nad besediščem, ki se oblikoslovno obnaša primerljivo, kar je izrednega pomena za jezikovno didaktiko – na ravni usvajanja slovenščine kot prvega in tudi drugega oz. tujega jezika. Priložnost za želeno dopolnitev leksikona je prinesel projekt 'Nova slovnica sodobne standardne slovenščine: viri in metode',[3] ki ima med cilji tudi razvoj metodologije za slovnični opis slovenščine na ravni oblikoslovja in besedotvorja. Namen prispevka je predstaviti prve projektne rezultate: metodologijo luščenja in ročnega urejanja oblikoslovnih vzorcev za samostalnike, pri čemer je pozornost usmerjena tudi v identifikacijo nalog za vsebinske izboljšave Sloleksa.

2 STROJNA PREDPRIPRAVA PODATKOV

2.1 Metodološko izhodišče

Nabor oblikoslovnih vzorcev za slovenščino, ki ga predstavlja prispevek, temelji na podatkih Sloleksa samega. Iz baze leksikona so s pomočjo v te namene pripravljenega programa pridobljeni kandidati za oblikoslovne vzorce, skupaj s

[3] Spletna stran projekta: http://slovnica.ijs.si/. Projekt (J6-8256) finančno podpira ARRS (2017–2020), vodja je Simon Krek.

pripadajočim besediščem in opredeljenimi razločevalnimi lastnostmi (oblikoskladenjske oznake in spremenljivi deli besednih oblik). Metoda temelji na smernicah (Dobrovoljc in dr. 2015b: 95–99), ki opredeljujejo tri glavne kriterije za kategorizacijo: (a) vzorci morajo biti strojno berljivi, (b) za optimalno procesiranje pisnega jezika je smiselno ločevati oblikoslovno in naglasno raven in (c) pristop mora temeljiti na jezikovni rabi. Kot je utemeljeno v navedenih smernicah, je za učinkovito strojno obravnavo k vzorcem nujno pristopiti formalistično, z identifikacijo razločevalnih značilnosti iz gradiva samega. Šele v drugem koraku nastopi jezikoslovna obravnava, pri kateri upoštevamo jezikovnosistemske značilnosti. Kot bo razvidno v nadaljevanju (pogl. 5.3), postavi vidik strojnega procesiranja v ospredje druga vprašanja kot jezikovnosistemski pristop, vendar prav sprememba zornega kota prinaša številne novosti, ki jih je mogoče uporabiti tudi za izboljšavo jezikovnega opisa.

Metoda kljub strojnim izhodiščem ni jezikoslovno nepodprta, saj Sloleks temelji na ročnih vnosih, ki na eni strani upoštevajo referenčne priročnike za slovenščino, na drugi pa odstope od jezikovnega standarda, ki so bili prepoznani pri razvoju slovničnega pregledovalnika Besana.[4] Strojno razvrščeni rezultati, še bolj pa primeri, ki ostanejo po procesu nerazvrščeni, nakazujejo mesta, ki se jim je pri prihodnjem razvoju leksikona treba posebej posvetiti. Za predstavljeno delo je zato ključen ročni pregled strojno pridobljenega gradiva, ki na primeru samostalnikov[5] postavlja načela za obravnavo ostalih pregibnih besednih vrst, opredeljuje nadaljnji razvoj programa za luščenje in ob razumevanju sestave leksikona opiše njegova šibka mesta ter korake za njegovo nadgradnjo (pogl. 5.1 in 5.2).

[4] Ta (kot tudi Sloleks) temelji na leksikalni zbirki Ases, ki je predstavljena v (Arhar, Holozan 2009). Sama metodologija ročnih vnosov v bazo, ki je ena od osnov za Sloleks, v literaturi še ni bila podrobneje predstavljena, jo pa na kratko povzemajo (Dobrovoljc in dr. 2015b).

[5] Samostalniki so bili izbrani za izhodišče, ker so v leksikonu Sloleks najpogosteje zastopana besedna vrsta – po podatkih iz (Dobrovoljc in dr. 2015b: 84) predstavljajo slabih 54 % iztočnic.

Vnaprej je treba opozoriti, da v prispevku navedeni seznam ni dokončen. Šele luščenje novega gradiva iz referenčnega korpusa bo omogočilo odločitve, ki jih zgolj na osnovi leksikonskega gradiva ni mogoče dokončno sprejeti. Rezultate je torej treba videti kot korak v razvojnem delotoku – oblikovanje izhodiščnega nabora za luščenja podatkov, s katerimi je nato mogoče nabor nadgrajevati.

2.2 Pridobivanje vzorcev iz leksikona

Pri strojnem pridobivanju vzorcev smo izhajali iz seznama lem, ki so zabeležene v leksikonu, natančneje skupkov leme in oznake za besedno vrsto, npr. *korak_S*, s čimer smo ločili besednovrstno raznolike enakopisne leme (npr. *lev_S* in *lev_P*). V prvem koraku smo za vsakega od tovrstnih skupkov iz leksikona izluščili vse zabeležene besedne oblike in njihove oblikoskladenjske oznake.[6] Ker v tabelaričnem formatu leksikona Sloleks oblike niso vedno razvrščene v predvidenem vrstnem redu (npr. ednina, dvojina, množina) oz. se red med različnimi lemami zaradi različnega števila oblik lahko razlikuje, smo v drugem koraku seznam oblik razvrstili glede na kanonični vrstni red oblikoskladenjskih oznak (pri glagolih npr. po osebi od prve do tretje in po številu od ednine do množine, na koncu še velelnik in neosebne glagolske oblike). Primer za samostalnik prikazuje Tabela 1.

Lema *čolnar*	*Ednina*	*Dvojina*	*Množina*
Oznake in oblike	Somei: čolnar	Somdi: čolnarja	Sommi: čolnarji
	Somer: čolnarja	Somdr: čolnarjev	Sommr: čolnarjev
	Somed: čolnarju	Somdd: čolnarjema	Sommd: čolnarjem
	Sometd: čolnarja	Somdt: čolnarja	Sommt: čolnarje
	Somem: čolnarju	Somdm: čolnarjih	Sommm: čolnarjih
	Someo: čolnarjem	Somdo: čolnarjema	Sommo: čolnarji

[6] Oblikoskladenjske oznake sistema JOS navajava s predpostavko, da njihovo pojasnjevanje ni potrebno. O sistemu označevanja je mogoče več prebrati na http://nl.ijs.si/jos/msd/html-sl/index.html ter v (Erjavec, Krek 2008).

Tabela 1: Razvrščene besedne oblike kot podlaga za identifikacijo vzorca.

V tretjem koraku smo za referenčno točko vzeli najkrajšo (oz. prvo najkrajšo) besedno obliko in jo strojno primerjali z vsemi ostalimi oblikami v seznamu, s čimer smo identificirali nespremenljivi del besede, ki je vsem oblikam skupen. V Tabeli 2 so spremenljivi deli obarvani rdeče.

Lema *čolnar*	*Ednina*	*Dvojina*	*Množina*
Oznake in oblike	Somei: čolnar	Somdi: čolnarja	Sommi: čolnarji
	Somer: čolnarja	Somdr: čolnarjev	Sommr: čolnarjev
	Somed: čolnarju	Somdd: čolnarjema	Sommd: čolnarjem
	Sometd: čolnarja	Somdt: čolnarja	Sommt: čolnarje
	Somem: čolnarju	Somdm: čolnarjih	Sommm: čolnarjih
	Someo: čolnarjem	Somdo: čolnarjema	Sommo: čolnarji

Tabela 2: Identifikacija nespremenljivega in spremenljivega dela besednih oblik.

V zadnjem koraku smo vsem oblikam odstranili nespremenljivi del besede in tako pridobili zaporedje spremenljivih delov. Tabela 3 prikazuje vzorec za lemo *čolnar*, pod isti vzorec pa spada še 1.152 občnoimenskih lem, npr. *direktor, davkar*, *enoceličar* in *guverner*.

Primer vzorca	*Ednina*	*Dvojina*	*Množina*
Oznake in spremenljivi deli oblik	Somei: -Ø	Somdi: -ja	Sommi: -ji
	Somer: -ja	Somdr: -jev	Sommr: -jev
	Somed: -ju	Somdd: -jema	Sommd: -jem
	Sometd: -ja	Somdt: -ja	Sommt: -je
	Somem: -ju	Somdm: -jih	Sommm: -jih
	Someo: -jem	Somdo: -jema	Sommo: -ji

Tabela 3: Strojno pridobljen pregibni vzorec za lemo *čolnar*.

Pri luščenju smo upoštevali tudi oblike, ki so bile v leksikonu označene kot nestandardne (npr. *hči* v tožilniku ednine). Kot take smo jih označili tudi v končnem izpisu pregibnega vzorca, in sicer tako, da smo jim pripisali znak #.

Tako nestandardne kot standardne variantne oblike smo ločili z znakom |. Primer podatkov za lemo *hči* prikazuje Tabela 4.

Vzorec za lemo *hči*	*Ednina*	*Dvojina*	*Množina*
Oznake in spremenljivi deli oblik	Sozei: -era_# | -er_# | -i	Sozdi: -eri	Sozmi: -ere
	Sozer: -ere	Sozdr: -era | -er	Sozmr: -era | -er
	Sozed: -eri	Sozdd: -erama	Sozmd: -eram
	Sozet: -er | -ero_# | -i_#	Sozdt: -eri	Sozmt: -ere
	Sozem: -eri	Sozdm: -erah	Sozmm: -erah
	Sozeo: -erjo	Sozdo: -erama	Sozmo: -erami

Tabela 4: Strojno izluščen vzorec za lemo *hči* z variantnimi (standardnimi in nestandardnimi) oblikami.

Vsakemu tako pridobljenemu vzorcu smo pripisali identifikacijsko številko ter seznam vseh lem, ki mu pripadajo. Končni izpis strojno izluščenih vzorcev je bilo tako mogoče razvrščati po produktivnosti (tj. številu lem, ki se pregibajo po določenem vzorcu), po besedni vrsti in po vsebnosti nestandardnih (#) ali variantnih (|) prvin.

2.3 Pridobivanje tipskega primera

Kot tipski primer oz. zgled za posamezni vzorec smo strojno izvozili podatek o pogostosti posameznih lem v korpusu Gigafida in znotraj vzorca poiskali tisto z najvišjo absolutno pogostostjo. Ta sicer preprosti postopek za veliko večino vzorcev prinese dobre rezultate, tj. splošno in predvidoma razumljivo besedišče, ki ustrezno reprezentira leme določenega vzorca, npr. *čas, predsednik, delo, življenje, država, možnost* itd. Samo v sedmih primerih smo za tipski primer izbrali lemo, ki ni bila prva po pogostosti. Razlogi so bili, da: (a) predlagani primer ni bil skladen s pisnim standardom (*studijo* zamenjamo s *ponču*); (b) predlagani primer ni bil intuitivno skladen z opisom vzorca (*živelj*, ki naj bi predstavljal enega od vzorcev za nežive samostalnike, zamenjamo z *žajbelj*); (c) predlagani primer (glede na subjektivne ocene) ni bil dovolj besednovrstno asociativen (*hvala* zamenjamo z *nafta*); (č) pri predlaganem

primeru so obstajali dvomi glede obstoja alternativne možnosti pregibanja (*Maze – Mazeja/Mazeta* zamenjamo z *Brezigar*); ali (d) je bil predlagani primer kako drugače sporen (*mami,* ki glede na Sloleks obstaja samo v ednini, zamenjamo z *madame*).[7]

3 JEZIKOSLOVNA ANALIZA IN UREJANJE V VZORCE

3.1 Namen in domet urejanja

Namen ročnega urejanja strojno pridobljenih podatkov je: (a) identificirati sistemsko in v rabi utemeljene vzorce in jih ločiti od rezultatov, ki se pojavljajo spričo šuma pri strojnem luščenju in nedoslednosti v leksikonu Sloleks; (b) urediti vzorce v skupine glede na vsebovanost in sorodnost podatkov; (c) poiskati in natančneje opredeliti variantnost, tako pri standardnih kot nestandardnih oblikah; (č) začrtati korake za nadaljnji razvoj programa in nadgradnje leksikona. Naloga, ki bo opravljena po prerazvrščanju podatkov, pripravi vzorcev za ostale besedne vrste in dodatnem luščenju iz korpusa Gigafida, bo vključevala doslednejše hierarhično urejanje pridobljenih skupin in dokončno poimenovanje posameznih vzorcev.

Analiza je potekala v tabelarični obliki. Za vsak vzorec so bili stolpično uvoženi spremenljivi deli vzorca skupaj z oblikoskladenjskimi oznakami, dodan je bil podatek o številu lem, ki ustrezajo vzorcu, njihov izpis ter tipski primer za vzorec. Ob ročnem pregledu rezultatov je bil v tabelo pripisan kratek opis vzorca z informacijami iz referenčnih jezikovnih priročnikov, ki dodatno pojasnjujejo posamezne dileme ali odločitve.[8] Primer urejenih podatkov za eno izmed

[7] Družbeno občutljivost jezikovnih priročnikov (o izzivih trenutnega stanja piše npr. Gorjanc 2017) je nujno zagotavljati na vseh stopnjah njihove priprave. Pri navedenem vzorcu sicer med petimi zajetimi lemami ni bilo najti dobrega tipskega primera in verjetno je, da bo v nadaljevanju vse gradivo premeščeno med primere, ki se pregibajo v vseh slovničnih številih.
[8] Glavni vir je bila Slovenska slovnica (Toporišič 2004), pri posameznih dilemah pa so bili podatki preverjeni tudi v slovarju SSKJ2 (Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja, www.fran.si, dostop oktober 2018).

skupin samostalnikov ženskega spola je v Tabeli 5. Na razlike med vzorci opozarjajo siva polja v tabeli.

Vzorec	SZ-1	SZ-1-ednina	SZ1-množina	SZ-1-j
Tip	***država***	***hvala***	***finance***	***alinea***
Število lem	12.092	865	58	15
Leme	abdikacija, abdukcija, abeceda, abecednica …	aerobika, aerodinamika, aerostatika, afrikanistika …	atmosferilije, bakanalije, bermudke, bikinke …	alinea, aloa, boa, gloria, goa …
Oblike	Sozei: -a	Sozei: -a		Sozei: -a
	Sozer: -e	Sozer: -e		Sozer: -e
	Sozed: -i	Sozed: -i		Sozed: -i
	Sozet: -o	Sozet: -o		Sozet: -o
	Sozem: -i	Sozem: -i		Sozem: -i
	Sozeo: -o	Sozeo: -o		Sozeo: -o
	Sozdi: -i			Sozdi: -i
	Sozdr: -Ø			Sozdr: -j
	Sozdd: -ama			Sozdd: -ama
	Sozdt: -i			Sozdt: -i
	Sozdm: -ah			Sozdm: -ah
	Sozdo: -ama			Sozdo: -ama
	Sozmi: -e		Sozmi: -e	Sozmi: -e
	Sozmr: -Ø		Sozmr: -Ø	Sozmr: -j
	Sozmd: -am		Sozmd: -am	Sozmd: -am
	Sozmt: -e		Sozmt: -e	Sozmt: -e
	Sozmm: -ah		Sozmm: -ah	Sozmm: -ah
	Sozmo: -ami		Sozmo: -ami	Sozmo: -ami
Opis	Osnovni vzorec za ženske samost. na -*a*, v rod. -*e*., prim. 1. ženska sklanj. (SS 289).	Podvzorec SZ-1 za edn. –verjetno legitimna raba dv. in mn. (preveriti korpus). Tipski primer ni najbolj intuitiven.	Podvzorec SZ-1 za mn.	Vzorec SZ-1 za primere, ko se beseda konča na zev (SS str. 290 umešča pod premene osnove).

Tabela 5: Ročno razvrščanje strojno pridobljenih vzorcev**.**

3.2 Načela urejanja vzorcev

Pri razvrščanju vzorcev so bila oblikovana in uporabljena naslednja načela:[9]

a) **Načelo nepopravljanja:** Kadar je strojno pridobljeni rezultat posledica leksikonskih pomanjkljivosti ali težav luščenja, se v tabeli rezultatov posebej označi, da gre za problem, sami podatki v leksikonu pa se v tem koraku še ne popravljajo. Glede na vrsto težave določimo, kakšna vrsta rešitve je potrebna v naslednjem koraku. Primere prinaša poglavje 5.2.

b) **Načelo vsebovanosti:** Vzorce, ki so omejeni na posamezno slovnično število, umestimo kot podvzorec ustrezajočih celotnih vzorcev. Primer kaže Tabela 5, kjer tip *država* predstavlja celovito deblo, tip *hvala* navaja samo edninske in tip *finance* samo množinske oblike.[10]

c) **Načelo sorodnosti:** Če se določen vzorec od drugega v jeziku pogostega vzorca loči v posamezni značilnosti, ki je jasno določljiva in omejena na posamezno obliko (pri čemer pa ne gre za variantnost), ga umestimo v isto skupino. Primer je umestitev tipa *alinea* ob tip *država* (Tabela 5). Načelo sorodnosti tipično uporabljamo za primere, kjer sta vzorca v pregibnih oblikah prekrivna, ločuje pa ju končnica leme (npr. umestitev tipa *dodo* ali *kamikaze* v skupino SM-1, gl. Tabela 7). Nastanek skupin je sekundaren v procesu urejanja: najprej se določi relevantnost posameznih strojno pridobljenih vzorcev, skupina pa nastane, ko so identificirani sorodni vzorci.

---

[9] Pri tem uporabljamo naslednja poimenovanja: *vzorec* je nabor oblikoskladenjskih oznak in spremenljivih delov oblik (kot prikazuje Tabela 3); *podvzorec* je delček drugega, obstoječega vzorca, ki pa je omejen (npr. samo na množino); *skupina* je nabor vzorcev, ki so si po značilnostih zelo podobni oz. sorodni, razlike med njimi pa so jasno opredeljene in zamejene; *izjema* je vzorec, za katerega je mogoče z gotovostjo predvidevati, da je zelo omejen in v sodobnem jeziku ni produktiven (npr. *otrok*).

[10] V podatkih se pojavljajo edninski in množinski podvzorci (ne pa tudi druge možne različice oz. kombinacije), v redkih primerih še podvzorci, ki pokrivajo posamezne oblike (npr. tip *EUR*, ki se sklanja z ničtimi končnicami, ima beleženo obliko samo za imenovalnik ednine).

d) **Načelo pogostosti:** V jeziku zelo pogoste in tipične značilnosti vedno povzročijo delitev na ločene vzorce. Po tem načelu stopijo v ospredje kot razločevalne denimo kategorija živosti pri moških samostalnikih, preglašenost končnic pri moškem in srednjem spolu ter izpuščanje polglasnika osnovne oblike oz. vrivanje samoglasnika pri vseh spolih. Načelo je pomembno, ker se tradicionalno te značilnosti obravavnavajo kot premene osnovnih paradigm, kar vpliva na strukturiranost in vsebino jezikovnega opisa. Razlike, ki jih (tudi za opis) prinese načelo pogostosti, natančneje prikazuje poglavje 5.3.

e) **Načelo produktivnosti:** Kot izjeme opredelimo vzorce, ki so izrazito omejeni na posamezne besede, obenem pa je mogoče predvideti, da v sodobnem jeziku niso produktivni (npr. vzorec za samostalnika *otrok* ali *kri*).[11] Če je vzorec potencialno produktiven, ga ohranimo v naboru, četudi ima nizko zastopanost (npr. (pod)vzorci, ki pokrivajo posamezne samostalnike *mozeg, bezeg* in *mezeg*).[12]

f) **Načelo specifičnosti:** Pri urejanju vzorcev trenutno ne uporabljamo združevalnih metakategorij. Primer je skupina SM-3 (Tabela 7), v kateri so zbrani vzorci z nepreglašenimi oblikami, pri katerih se v deblu izpušča polglasnik (*meter, posel, kamen* ipd.). Soglasnik, ki se posledično pojavi v pregibnem delu oblike (*r, l, n*), glede na dano metodologijo opredeljuje in ločuje vzorec od drugih. Generalizirana oblika (te je mogoče vpeljati naknadno) bi vključevala metakategorijo za nabor soglasnikov in posamezne vzorce združila v enega.[13]

g) **Načelo enovitosti:** Kadar je izluščeni vzorec sestavljen iz več

[11] Z določeno mero prizanesljivosti, npr. primeri *uho, oko, igo* so umeščeni med vzorce, ne izjeme, vsaj dokler ne izvedemo preverbe zastopanosti vzorca v referenčnem korpusu. Preverba podatkov v korpusu bo pomagala jasneje ločiti kategoriji vzorcev in izjem – če se bo njuno ločevanje za pripravo končnega rezultata sploh potrdilo kot smiselno oz. potrebno.

[12] V isti vzorec kot *mezeg* bi npr. spadalo lastno ime *Drozeg,* ki v Sloleks ni uvrščeno (v korpusu Gigafida se pojavi v dveh konkordancah).

[13] 26 (pod)vzorcev v skupini SM-3 je na tak način mogoče strniti v 2.

posameznih identificiranih vzorcev, ga označimo s posebno oznako, ki nakazuje, da je v nadaljnjem procesiranju besedo treba umestiti k vsem ustrezajočim vzorcem (več v poglavju 3.3).

h) **Načelo omejene variantnosti:** Variantnost beležimo, kadar se izbira pojavlja na ravni posameznih oblik, pri čemer so vse vrste tovrstnih pojavitev jasno opredeljene (Tabela 6). V redkih primerih, kjer je posebnosti pri pregibanju veliko, obenem pa se pojavlja sum, da podatki o variantnosti v leksikonu niso dosledno pripisani ali ažurni, so primeri označeni za nadaljnje analize in niso dodani med vzorce (npr. *drog, zid, voz, vol*).

i) **Načelo ločenih vnosov:** Kadar se dvojnica v leksikonu pojavlja na ravni imenovalnika ednine (npr. *penal/penale*), se predlaga leksikonski vnos ločenih lem.

j) **Načelo standardnosti:** Variantnosti, ki so posledica nestandardnih oblik (npr. *paparacom/paparacem*), pri pripravi vzorcev zanemarimo. Nestandardne variante beležimo, da jih bo v nadaljevanju mogoče sistematično preveriti v korpusnem gradivu (gl. tudi poglavje 3.3).

3.3 Standardne variante ter dvojni vzorci

Kadar se določena lema glede na Sloleks pregiba po več vzorcih, se po trenutni metodologiji podatki za vse vzorce izpišejo združeno. Pri analizi takšne primere ročno označimo z namenom, da bodo v naslednjem koraku ustrezno strojno prerazvrščeni in bo posledično tovrstnim lemam pripisanih po več ločenih vzorcev. Analiza je pokazala, da se dvojni vzorci pojavljajo pri naslednjih skupinah podatkov: (a) samostalniki, ki se lahko (glede na pomen) pregibajo po paradigmi za živo ali neživo (npr. *tip, nosilec, dvojček; Anton, Diego*); (b) samostalniki, ki se lahko pregibajo s podaljševanjem osnove ali brez (npr. *glas*[*ov*]*i, svet*[*ov*]*i, mand[e]lj[n]a, okvir*[*j*]a, *premier*[*j*]a); (c) samostalniki, kjer se v zapisu upošteva preglas ali ne (npr. *radiom/radiem*); (č) samostalniki, kjer se lahko polglasnik izpušča ali ne (npr. *meseca/mesca*); (d) samostalniki, ki se lahko pregibajo po paradigmah za različne spole (npr. *DNK-ja* ali *DNK-Ø,*

ledvica, skripta); (e) priimki, ki izkazujejo sklanjatev za moški in ženski spol (*Potočnik – Potočnika* vs. *Potočnik – Potočnik*); (f) lastna imena, ki so lahko različnih vrst, npr. ime ali priimek, osebno ali zemljepisno ime (npr. *Miro*, *Chelsea, Pearl)*; (g) primeri, kjer lahko pride do različnega podaljševanja osnove (npr. *Arne – Arneta/Arneja*). Pri analizi samostalnikov, ki se lahko pregibajo po različnih vzorcih, so se razkrile tudi določene leksikonske nedoslednosti, h katerim se vračamo v poglavju 5.2.

Oblikovne variante pri pripravi vzorcev beležimo kot (neobvezni) dodatni del v imenu katerega koli obstoječega vzorca. Tako se npr. vzorec za tip *gospodje* loči od vzorca za tip *predsednik* po dodani oznaki za varianto V1 (SM-1(ž) vs. SM-1(ž)-V1, gl. Tabelo 7). Identificirani nabor variant navaja Tabela 6. Tudi na ravni beleženja variant se kažejo številna mesta leksikona, pri katerih bi bilo mogoče podatke urediti in poenotiti, predvsem pa je nujno vključiti preverbo obstoja v korpusnih podatkih in ločiti jezikovnosistemsko utemeljene možnosti od tistih, ki so prisotne v sodobni jezikovni rabi.[14]

Varianta	**Opis**	**Primer**
Moški spol V1	Imenovalnik množine: *-(ov)i* | *-je*	*gospodi/gospodje*
Moški spol V2	Rodilnik ednine, pri samostalnikih, ki izražajo živost, tudi tožilnik ednine: *-a* | *-u*	*mira/miru*
Moški spol V3	Mestnik množine, v določenih primerih tudi dvojine: *-eh* | *-(ov)ih*	*gostih/gosteh*

[14] Podatki o arhaičnih in stilnih variantah so lahko dragoceni za določene naloge obdelave naravnega jezika in jih nikakor ne gre zanemariti, za naloge, ki so vezane na procesiranje sodobnega in splošnega (če je mogoče tako imenovati jezik, ki ga reprezentira pisni referenčni korpus) jezika pa lahko njihova vsebnost deluje kontraproduktivno. Veljalo bi torej posebej označevati dvojnice, ki v sodobni rabi nimajo več potrditve, obenem pa leksikonske informacije (ustrezno opremljeno z metapodatki) obogatiti tudi s podatki iz korpusov, kot sta IMP, korpus starejših slovenskih besedil (Erjavec 2015), in Janes, korpus računalniško posredovane komunikacije (Fišer in dr. 2016). Slednji je nepogrešljiv vir tudi za nadaljnjo obravnavo nestandardnih oblik.

Moški spol V4	Orodnik množine: -(*ov*)*i* | *-mi*	*mostovi/mostmi*
Ženski spol V1	Rodilnik dvojine in množine: *-Ø* | *-a*	*vod/voda*
Ženski spol V2	Rodilnik dvojine in množine: *-ac* | *-c*	*ovc/ovac*
Ženski spol V3	Orodnik ednine: *-ijo* | *-jo*	*rebrjo/rebrijo*
Srednji spol V1	mestnik dvojine in množine: *-eh* | *-ih*	*sencih*/*senceh*

Tabela 6: Oblikoslovne variante, ki se pojavljajo v identificiranih vzorcih**.**

3.4 Nestandardne variante

Kot nestandardne dvojnice se v leksikonu beležijo pogostejše težave jezikovne rabe. Pri obravnavi samostalnikov je najti primere: (a) nestandardnega sklanjanja kratic brez vezaja (npr. *CDja*); (b) neustreznega ne/izpuščanja polglasnika na ravni oblike ali leme (npr. *filem, ansambl, vrteca, luknj, podlaht*); (c) neustreznega ne/podaljševanja osnove pri pregibanju (npr. *filterja*); (č) neustreznega ne/preglaševanja (npr. *paparacom*); (d) neustreznega tvorjenja osnovne oblike po analogiji s pregibnimi oblikami (npr. *bukva*); in (e) primere regionalnih oblik (npr. *v Prekmurji*). K možnim izboljšavam beleženja nestandardnih oblik se vračamo v poglavju 5.2.

4 REZULTATI

Nabor vzorcev navajava v tabelarni obliki, ločeno za vse tri spole. Občnoimenski in lastnoimenski podatki so predstavljeni skupaj glede na vzorec, ločuje jih podpičje. Trenutno so pri lastnoimenskih samostalnikih moškega spola ločeno navedeni vzorci, ki pokrivajo priimke, in sicer zato, ker se slednji v leksikonu vedno pojavljajo tudi z vzorcem za sklanjanje z ničto končnico za ženski spol. Vzorci so razvrščeni po skupinah, ki so kratko opisane. Za vsak vzorec je navedena (trenutna) kratka koda ter opredelitev, ali je vzorec v leksikonu opredeljen v celoti ali le delno (npr. za ednino ali množino). Sledi navedba

števila lem, ki jih vzorec v leksikonu pokriva,[15] ter strojno pridobljenih tipskih primerov. V prispevku ni prostora za navajanje celotnih vzorcev (kakor so denimo prikazani v Tabeli 3), vendar je do vseh podrobnosti mogoče dostopati s pomočjo spletne različice leksikona (http://www.slovenscina.eu/sloleks).

Koda	**Celovitost**	**Št. lem (O; L)**	**Tipski primer(i)**
1. skupina: Osnovni nepreglašeni vzorci za neživo (n) in živo (ž); posebej sta vzorca za leme na *-o* in *-e* . Variante razlaga Tabela 6.			
SM-1(n)	celotna	4.878; 2	*čas; Windows*
	ednina	523; 444	*promet; Maribor*
SM-1(n)-V1	celotna	1	*ud*
SM-1(n)-V2	celotna	4	*mir*
	ednina	6	*sram*
SM-1(n)-V2, V3	celotna	2	*nos*
SM-1(n)-V3, V4	celotna	1	*kol*
SM-1(ž)	celotna	2.528; 1.122; 434	*predsednik; Potočnik; Janez*
SM-1(ž)-V1	celotna	32	*gospod*
SM-1(ž)-V1, V3	celotna	1	*gost*
SM-1	množina[16]	11; 25	*hemoroidi; Helsinki*
SM-1o(n)	celotna	98; 2	*evro; Yugo*
	ednina	18; 42	*vaterpolo; Nato*
SM-1o(ž)	celotna	25; 125; 75	*dodo; Branko; Šukalo*
SM-1e(n)	celotna	9	*polfinale*
	ednina	1; 7	*pasodoble; Google*
SM-1e(ž)	celotna	1; 21; 8	*kamikaze; Stone; Mike*

[15] Podatki o številčni zastopanosti se bodo v končni različici povečali na račun prerazvrščenih samostalnikov. Z dopolnjevanjem skupin in popravljanjem nedoslednosti v korpusu se lahko v končni različici spremeni tudi izbira tipskega primera (npr. *konj* namesto redkega *prakonj*).
[16] Množinske oblike ne izkazujejo razlik na ravni živosti, zato jih po tem kriteriju strojno ni mogoče ločevati.

Izjeme: *SM-1(ž)-otrok*, *SM-1(ž)-človek*, *SM-1(n)-las*.			
2. skupina: Osnovni preglašeni vzorec za neživo (n) in živo (ž); posebej je naveden vzorec za leme na *-o*.[17]			
SM-2(ž)	celotna	853; 790; 53	*prijatelj; Majdič; Franc*
SM-2(n)	celotna ednina	640 124; 69	*razvoj* *hokej; Kranj*
SM-2	množina	5; 34	*tisoči; Radenci*
SM-2o(ž)	celotna	7	*Franjo*
SM-2o(n)	celotna	7	*poncho*
Izjeme: *SM-2(ž)-prakonj*, *SM-2(ž)-mož*.			
3. skupina: Nepreglašeni vzorec za neživo (n) in živo (ž), pri katerem se izpusti polglasnik. Deljeni so glede na soglasnik, ki se zato pojavi v spremenljivem delu.			
SM-3k(n)	celotna ednina	829 6; 5	*odstotek* *nameček; Podčetrtek*
SM-3k(ž)	celotna	232; 113; 15	*deček; Lipovšek; Božiček*
SM-3m(n)	celotna ednina	264 207; 1	*sejem* *turizem; Videm*
SM-3m(ž)	celotna	1; 1	*mikroorganizem; Erazem*
SM-3r(n)	celotna ednina	151 9; 6	*meter* *koper; Koper*
SM-3r(ž)	celotna	41; 26	*minister; Bešter*
SM-3ar(ž)	celotna	2	*Aleksandar*
SM-3ar(n)	ednina	2	*Zadar*
SM-3l(n)	celotna ednina	41 7	*posel* *Basel*
SM-3l(ž)	celotna	7; 7; 2	*osel; Rupel; Pavel*
SM-3ll(ž)	celotna	2	*Rusell*
SM-3n(n)	celotna ednina	29 2; 14	*kamen* *česen; München*
SM-3n(ž)	celotna	3; 23; 2	*oven; Verboten; Domen*

[17] Različica za leme na *-e* ni izpričana, čeprav jo je mogoče predvideti za imena tipa *Djordje* (651 pojavitev v korpusu Gigafida). Kot rečeno, bo luščenje podatkov iz korpusa omogočilo identifikacijo in dopolnitev manjkajočih vzorcev.

SM-3g(n)	celotna ednina	1 1	*mozeg* *bezeg*
SM-3g(ž)	celotna	1	*mezeg*
SM-3t(n)	celotna ednina	1 1	*hrbet* *ocet*
SM-3t(ž)	celotna	1	*valpet*
SM-3s(ž)	celotna ednina	1 1	*pes* *oves*
Izjeme: *SM-3r(n)-veter*; *SM-3r(n)-blagor*.			
4. skupina: Preglašeni vzorec za neživo (n) in živo (ž), pri kateri se izpusti polglasnik. Deljeni so glede na soglasnik, ki se zato pojavi v spremenljivem delu.			
SM-4c(n) SM-4c(n)-V3	celotna ednina celotna	405 24; 18 1	*marec* *svinec; Gradec* *konec*
SM-4c(ž)	celotna	1.906; 98; 65	*igralec; Mavec; Avstrijec*
SM-4ac(ž)	celotna ednina	9 2	*Badovinac* *Karlovac*
SM-4lj(n)	celotna ednina	52 3; 7	*čevelj* *žajbelj; Bruselj*
SM-4lj(ž)	celotna	10; 43	*rabelj; Avbelj*
SM-4nj(n)	celotna ednina	12 1	*ogenj* *Sovodenj*
SM-4nj(ž)	celotna	3	*suženj*
SM-4š(n)	ednina	1	*Mengeš*
5. skupina: Vzorec za sklanjanje z uporabo vezaja, pri čemer se uporabljajo preglašene (*-ja*) in nepreglašene (*-a*) končnice.			
SM-5ja(n)	celotna ednina	62 66; 357	*m* *DDV; BMW*
SM-5a(n)	ednina	4	*GSM* (tudi po SM-5ja(n)-ednina)
6. skupina: Vzorec za sklanjanje z ničtimi končnicami.			
SM-6	celotna ednina 'Somei' 'Sometn'	3 15; 131 53 1	*mio* *foto; New* *EUR* *poštev*

7. skupina: Vzorec za leme na *-a* ali *-ja,* ki je enak ženskim vzorcem, ali pregibanje po preglašenih in nepreglašenih vzorcih za moški spol.[18]			
SM-7ja(ž) ali SM-9ja(ž)	celotna	18; 7; 3	*zborovodja; Burja; Mitja*
SM-7a(ž) ali SM-1a(ž)	celotna	15; 54; 9	*panda; Slana; Miha*
SM-7a(ž) ali SM-2a(ž)	celotna	14; 32; 9	*kuža; Franca; Matija*
8. skupina: Vzorec za neživo (n) in živo (ž), ki v dvojini in množini izraža podaljšavo z *-ova-*. Variante razlaga Tabela 6.			
SM-8(n) SM-8(n)-V2 SM-8(n)-V2,V3, V4	celotna celotna celotna	6 11 1	*sok* *strah* *most*
SM-8(ž) SM-8(ž)-V2	celotna celotna	2 1	*bog* *tat*
9. skupina: Vzorec za neživo (n) in živo (ž), ki podajšuje osnovo z *-j*, *-t* ali *–n*. Variante razlaga Tabela 6.			
SM-9j(n)	celotna ednina	501 40; 65	*denar* *humor; Tivoli*
SM-9j(ž)	celotna	1.152; 507; 125	*direktor; Brezigar; Igor*
SM-9t(n)	celotna	1	*kofe*
SM-9t(ž) SM-9t(ž)-V1	celotna celotna	6; 80; 23 2	*pezde; Blagne; Jože* *oče*
SM-9n(n)	celotna	3	*buhtelj*
Izjeme: *SM-9-dan1* (*dan - dneva*) in *SM-9-dan2* (Somer: *dan - dne*).			
10. skupina: Vzorec, podoben pridevniškemu.			
SM-10(ž)	celotna	5; 14	*moški; Cetinski*

Tabela 7: Skupine vzorcev za samostalnike moškega spola.

[18] Vzorce, načeloma enake ženskim, beležimo tudi kot vzorce za moški spol. Enako ustrezna možnost bi bila pripis vzorcev za ženski spol ali nenazadnje oblikovanje poimenovanj tako, da spol samostalnika (ali tudi besedna vrsta, gl. SM-10(ž)) ni ločevalna značilnost. S tem bi zmanjšali število vzorcev, tudi denimo pri beleženju sklanjanja z ničtimi končnicami. Odločitev trenutno še ni jasna, mora pa biti optimalna z vidika strukturiranja podatkov v bazi.

Koda	Celovitost	Št. lem (O; L)	Tipski primer(i)
1. skupina: Osnovni vzorec za samostalnike ženskega spola, ki se končajo na *-a*. Ločeni so primeri, kjer se lema konča na zev. Variante razlaga Tabela 6.			
SZ-1	celotna ednina množina	12.092; 458 865; 492 58; 181	*država; Amerika* *nafta; Slovenija* *finance; Jesenice*
SZ-1-V1	celotna	9	*voda*
SZ-1j	celotna	15; 23	*alinea; Maria*
Izjeme: *SZ-1-gospa; SZ-1-Golte*			
2. skupina: Osnovni vzorec za samostalnike ženskega spola, ki se ne končajo na *-a*.			
SZ-2	celotna ednina množina	5.202 22; 5 4	*možnost* *last; Podpeč* *obresti*
3. skupina: Vzorec za leme, ki se končajo na *-ev*.			
SZ-3	celotna ednina množina	819 1 2	*odločitev* *Lokev* *Ponikve*
4. skupina: Vzorec za samostalnike ženskega spola, ki se ne končajo na *-a* in imajo v množini v določenih sklonih v spremenljivem delu oblike *-e-*.			
SZ-4	celotna množina	95 2; 2	*stran* *sani; Ravni*
Izjeme: *SZ-4-kri; SZ-4-Žiri*			
5. skupina: Vzorec za samostalnike ženskega spola na *-a*, kjer se v rodilniku dvojine in množine vriva *e* ali *i*, redko tudi *a*.			
SZ-5r SZ-5r-V1	celotna množina celotna	64; 6 2; 1 1	*igra; Petra* *citre; Pekre* *sestra*
SZ-5nj	celotna množina	53 4	*izkušnja* *Bitnje*
SZ-5lj SZ-5lj-V1	celotna množina celotna	34 2; 13 1	*kaplja* *grablje; Trbovlje* *zemlja*
SZ-5l SZ-5l-V1	celotna množina celotna	33 3; 1 2	*megla* *orgle; Murgle* *metla*
SZ-5m	celotna	28	*tekma*

SZ-5n	celotna množina	24; 2 3	*opna; Vesna* *Ravne*
SZ-5v	celotna	15	*spužva*
SZ-5j	celotna množina	5; 5 2; 4	*ladja; Katja* *škarje; Nazarje*
SZ-5c-v2	celotna	1	*ovca*
SZ-5k-v2	celotna množina	1 1	*deska* *nečke*
Izjeme: *SZ-5r-mati, SZ-5r-hči*			
6. skupina: Vzorec za sklanjanje z ničtimi končnicami.			
SZ-6	celotna ednina množina	7; 121 5 1; 1	*lady; Jennifer* *madame* *OI; ZDA*
7. skupina: Vzorec za samostalnike ženskega spola, ki se ne končajo na *-a* in vsebujejo izpustljiv polglasnik. V dv. in mn. so končnice z *-i-* (*boleznima*).			
SZ-7-en	celotna	12	*bolezen*
SZ-7-enj	celotna	2	*povodenj*
SZ-7-el	celotna množina	3 5	*misel* *jasli*
SZ-7-em	celotna	1	*pesem*
8. skupina: Vzorec za samostalnike ženskega spola, ki se ne končajo na *-a* in vsebujejo izpustljiv polglasnik. V dv. in mn. so končnice z *-e-* (*ravnema*).			
SZ-8-en-v3	celotna množina	2 2	*raven* *Ravni*
SZ-8-an-v3	celotna	1	*ravan*
SZ-8-er-v3	celotna	2	*reber*
SZ-8-et-v3	celotna	2	*lahet*
9. skupina: Vzorec, podoben pridevniškemu.			
SZ-9	celotna	2	*častita*

Tabela 8: Skupine vzorcev za samostalnike ženskega spola.

Koda	**Celovitost**	**Št. lem (O; L)**	**Tipski primer(i)**

1. skupina: Osnovni vzorec za preglašene samostalnike.			
SS-1	celotna ednina množina	5.740 197; 41 6	*življenje* *zdravje; Celje* *vratca*
SS-1-V1	celotna	1	*sence*
2. skupina: Osnovni vzorec za nepreglašene samostalnike.			
SS-2	celotna ednina množina	418 469; 17 15; 7	*delo* *mleko; Kosovo* *vrata; Selca*
Izjeme: *SS-2-Rova*			
3. skupina: Vzorec za preglašene samostalnike, kjer se v rodilniku dv. in mn. vriva -i- ali -e-.			
SS-3i	celotna	564	*podjetje*
SS-3e	celotna	5	*ozemlje*
4. skupina: Vzorec za nepreglašene samostalnike, kjer se v rodilniku dv. in mn. vriva -e-.			
SS-4v	celotna	443	*ministrstvo*
SS-4l	celotna	21	*geslo*
SS-4n	celotna	10	*okno*
SS-4r	celotna množina	7 2	*jutro* *jetra*
SS-4m	celotna	2	*pismo*
Izjeme: *SS-4l-tla; SS-4n-dno*			
5. skupina: Vzorec, ki vsebuje podaljševanje osnove s *-t-*, *-n-* ali *-s-*.			
SS-5t	celotna	16	*dekle*
SS-5n	celotna	11	*ime*
SS-5s	celotna	2	*oje*
6. skupina: Vzorec za samostalnike na -o, ki imajo podaljšavo z *-es-*.			
SS-6	celotna	9	*telo*
SS-6h	celotna	1	*uho*
SS-6k	celotna	1	*oko*
SS-6g	celotna	1	*igo*

Izjeme: *SS-6-črevo*			
7. skupina: Vzorec za sklanjanje z ničtimi končnicami.			
SS-7	'Sosei' in 'Soset'	3	*dopoldne*
8. skupina: Vzorec, podoben pridevniškemu.			
SS-8o	celotna ednina	9 72	*valentinovo* *Laško*
SS-8e	ednina	5	*Trebnje*

Tabela 9: Skupine vzorcev za samostalnike srednjega spola.

5 VREDNOST PODATKOV ZA NADALJNJI RAZVOJ VIROV

5.1 Program za pridobivanje vzorcev

Rezultati kažejo, da program za strojno pridobivanje vzorcev (pogl. 2.2) ponuja dobra izhodišča za nadaljnjo analizo. Možne so izboljšave, ki bodo optimizirale delo za preostale besedne vrste. Trenutno se denimo v podatkih skupaj izpisujejo enakopisni samostalniki različnih spolov (npr. *prst, čelo, bit, tv*), kar je že pri luščenju mogoče obravnavati ločeno, saj gre za ločene leksikonske enote. Kot drugo, težave so na ravni izpisa spremenljivega dela dvojnic: kadar sta varianti pri različnih besedah v leksikonu navedeni v različnem vrstnem redu (npr. *pand<u>ov</u> / pand* vs. *nadvojvod / nadvojvod<u>ov</u>*), program vzorca obravnava kot različna, kar bi bilo mogoče popraviti z dodatnim korakom, ki bi preverjal oz. zanemaril tovrstna zaporedja. Dodati bi bilo mogoče tudi predrazvrščanje delnih vzorcev pod celovite, kot tudi že izhodiščno umeščanje besed v dvojne vzorce, čeprav je pri slednjih primerih koristno, da se v prvem koraku izpisujejo ločeno in analizirajo ročno. Brez dvoma pa je strojno prerazvrščanje nujno razviti za drugi korak obravnave.

5.2 Leksikon Sloleks

Kot je bilo omenjeno na več mestih dosedanje razprave, rezultati pričujočega dela niso zanimivi samo za dodajanje novih leksikonskih informacij (vzorcev), ampak tudi za urejanje obstoječe vsebine leksikona. Razvrščanje samostalnikov

v vzorce in skupine namreč izpostavi primere, v katerih se pojavljajo neregularnosti. Kot stranski rezultat torej dobimo seznam lem, ki potrebujejo pregled. Skupine težav, ki jih razkriva analiza, so naslednje: (a) v trenutni različici Sloleksa je možnost podaljševanja osnov mestoma beležena nedosledno, npr. za razliko od vzorca *glas,* kjer se v dvojini in množini lahko pojavlja podaljšava -ov-, imajo *pas, sin, top* v vzorcu beleženo to možnost samo v dvojini (*pasa/pasova* vs. *pasovi*), *val* in *trak* pa samo v množini (*vala* vs. *valom/valovom*); (b) variantnost je včasih neidentificirana, npr. samostalnik *megla* nima pripisanih končniško naglašenih variant v rodilniku dvojine in množine (beležena je oblika *megel,* ne pa tudi *megla*); (c) kot omenjeno (pogl. 3.3) so težave z nedosledno beleženimi in v rabi neizpričanimi starinskimi ali stilno zaznamovanimi variantami, npr. beleženje dvojnice na *-eh* v mestniku množine moškega spola (*drogeh, zideh, noseh*), dvojnice na *-mi* v orodniku množine moškega spola (npr. *mostmi)* ali vrivanja samoglasnika *-a-* v rodilniku dvojine in množine ženskega spola (*desak*); (č) mestoma nedosledno je beleženje celovitosti pregibalnih možnosti, npr. ime *Karmen* ima v leksikonu samo edninski del vzorca, čeprav tip *Jennifer* prinaša celotni vzorec); (d) pri več podvzorcih, ki so omejeni na posamezno število, bi bilo mogoče predvideti tudi rabo drugih oblik (npr. tip *hokej, Slovenija, mami* imajo beležene samo oblike za ednino); (e) v Sloleksu se pojavlja navajanje variantnosti v imenovalniku ednine, ki bi zahtevalo ločene leksikonske vnose (npr. *penal/penale*); (e) sicer redko se pojavljajo napake pri beleženju kategorije živosti samostalnikov moškega spola (npr. *adagio* je umeščen v vzorec za živo); (f) redkejši so tudi lapsusi na ravni vpisa oblik, npr. pri samostalniku *počasnost* se pojavlja med oblikami *prepočasnost*, zaradi česar program izpiše neustrezen vzorec; (g) in nenazadnje se v podatkih pojavljajo določene problematične besednovrstne umestitve, npr. besede *jesti* med samostalnike srednjega spola[19]

[19] Umestitev sledi Slovenski slovnici (Toporišič 2004: 301), ki samostalniško rabljene nedoločnike omenja pri 3. srednji sklanjatvi, primer *dobro jesti in piti.* Besednovrstno gre takšne primere v leksikonu obravnavati kot glagole.

ter (h) vsaj v podatkih, ki so na voljo v vmesniku, primeri nestandardno črkovanih lem, pri katerih pa nestandardnost ni označena, npr. *jeterca, pluča*.

Nekatere od naštetih težav se pojavljajo sporadično in jih je treba obravnavati ročno, dober delež pa je mogoče urediti sistemsko. Preverba pojavljanja oblik v različnih številih in ne/uporabljenih podaljšav sta denimo nalogi, ki ju je mogoče v korpusu preveriti strojno za celoten nabor ustrezajočih lem. Predvideno je, da bo podatke, pridobljene s strojnimi luščenji za določene prepoznane probleme, treba natančneje analizirati, npr. rabo redkejših dvojnic, ki sovpadajo z v jeziku pogosto rabljenimi oblikami (npr. rodilnik množine *vodá*, ki je v zapisu prekriven z osnovno obliko *vôda*).[20]

Podatki o nestandardnih oblikah oz. tipičnih odstopih od trenutnega jezikovnega standarda na ravni oblikoslovja so dragoceni za uporabniško skupnost (Arhar Holdt in dr. 2013), vendar so trenutno v leksikonu beleženi zelo sporadično. V nadaljevanju je obstoj nestandardnih dvojnic treba preveriti sistematično pri vseh lemah, ki se pregibajo po določenem vzorcu. V leksikonu navedene nestandardne oblike vzamemo za izhodišče preverbe. Primer sta nestandardni obliki *bukva, bukvo* namesto standardne *bukev;* tovrstna nestandardnost je v leksikonu beležena pri dveh samostalnikih, mogoče pa jo je iskati za vse samostalnike, ki se sklanjajo po vzorcu SZ-3 (*odločitev*). Kot omenjeno, je za preverbo nestandardne morfologije smiselno vključiti korpus računalniško posredovane komunikacije Janes (Fišer in dr. 2016). V literaturi (Dobrovoljc in dr. 2015b: 100) je tudi že bilo izpostavljeno, da bi natančnejša kategorizacija (nestandardne in standardne) variantnosti omogočila naprednejše povezovanje leksikona z drugimi viri, npr. Slogovnim priročnikom (Krek in dr. 2013). Slednji je zasnovan na osnovi tipičnih uporabniških jezikovnih zadreg in kot tak ponuja informacije in rešitve, komplementarne leksikonskim podatkom, kot tudi slovničnemu opisu, ki se mu posvečamo v

[20] Pri analizah je treba imeti v mislih, da oblikoskladenjsko označevanje korpusa Gigafida temelji na podatkih leksikona Sloleks in posledično odraža zgoraj navedene pomanjkljivosti.

nadaljevanju.

5.3 Slovnični opis

Pridobljene podatke je mogoče uporabiti za dopolnitev obstoječega slovničnega opisa oz. razmislek, kako zasnovati slovnični opis , da bo tovrstne izsledke lahko izčrpno in informativno vključeval. Pristop iz gradiva je dragocen, ker prinaša podatke o pogostosti in s tem tipičnosti pomembnih razlikovalnih elementov med vzorci, na osnovi pogostosti izbrane tipske primere, izhodiščno pa zato, ker temelji na avtentičnih podatkih o sodobnem slovenskem jeziku v rabi. Vprašanje prihodnjega slovničnega opisa, ki je seveda kompleksno in si v slovenskem prostoru brez dvoma zasluži več pozornosti, presega domet prispevka; vseeno pa se zdi na tem mestu smiselno izpostaviti nekaj primerov za boljšo predstavo, kako naprej.

Podatke, ki smo jih z opisano metodo pridobili za samostalnike srednjega spola, primerjamo s podatki, ki so na voljo v Slovenski slovnici (Toporišič 2004: 297–301). Slovnično poglavje, ki obsega samo štiri strani in pol, je na vrhnjem nivoju strukturirano po sklanjatvah. Največ informacij je na voljo za prvo srednjo sklanjatev (vzorec *mesto*), začenši z (a) naborom končnic, sledijo: (b) po odstavkih opisane premene osnove (daljšanje osnove s *-t-, -n-* in *-s-;* kakovostne premene naglasov na *-e-* in *-o-* v osnovi besede, npr. *srêbru/srébru*; vrivanje *-e-* oziroma *-i-* v rod. dv. in mn. pri določenih primerih); (c) po odstavkih opisane premene končnic (v im. ed. ničta končnica pri besedah, ki podaljšujejo osnovo; preglas za *c, č, j, š, ž;* končnica *-eh* namesto *-ih* v mest. dv. oz. mn. (npr. *drveh*); množinski končnici *-emi*, *-mi* (npr. *drvmi*); množinski varianti *blaga-blagovi*) in na koncu (č) naglasi (najprej jakostni po naglasnih tipih, nato pa še tonemski po naglasnih tipih in akutiranih oz. cirkumflektiranih osnovah). V naslednjem podpoglavju o 2. srednji sklanjatvi izvemo, da je ni, 3. sklanjatev z ničto končnico je z nekaj primeri omenjena kot redka (*vremja*), 4. srednja sklanjatev (*Krško*) prinaša končnice za ednino in opombo, da sta množina in dvojina redki, vendar ne nemogoči.

Na drugi strani strojno luščenje podatke loči v osem skupin, v katerih je daleč najpogosteje izpričan tip *življenje,* torej vzorec s preglašenimi končnicami, ki ga slovnica ne navaja eksplicitno. Tip *delo* oz. *mesto,* ki je edini primer, pri katerem slovnica navaja celotni nabor končnic, je od tipa *življenje* glede na število vsebovanih lem skoraj 14-krat redkejši. Preglaševanje se prvič omeni šele med premenami končnic (in niti ne na prvem mestu, prehiti ga pregibanje besed s podaljševanjem osnove, ki so v seštevku od tipa *življenje* redkejše približno 198-krat). Čeprav je v jezikoslovnem smislu preglaševanje mogoče razumeti kot sekundaren pojav, se torej kaže potreba, da slovnični opis – še zlasti, če je slednji pripravljen tudi za jezikovnodidaktične namene – izhaja iz podatkov o sinhroni jezikovni rabi in vsebine predstavlja na način, da so tipične in pogoste značilnosti postavljene v ospredje. Pogostost je pomembna tudi pri drugih vzorcih. Npr. v tretji in četrti najpogostejši vzorec po strojnem luščenju umeščamo primere, kjer se v rodilniku dv. in mn. vrivata samoglasnika *-e-* ali *-i-*. Ti vzorci so v slovnici omenjeni pod premenami osnove, vendar na izjemno nepregleden in nesistematičen način (ibid: 298):

> Če se osnova samostalnikov končuje na nezvočnik – zvočnik ali na *rj, vj,* se v rod. mn./dv. pred (drugi) zvočnik vriva polglasnik, pred *j* pa *i*: *povesmo – povesem, kraljestvo – kraljestev, stêgno – stêgen, jétra – jéter, sêdlo – sêdel* proti *obzídje – obzídij, morje – morij, nedrje, nedrij, gorovje – gorovij*. Skupna imena na *-je* in s koncem podstave na *n* (*osten-je*), sklop *nj* prav tako razbijajo z *i: osténij*. Kadar občutka za tvorjenost ni, tudi ni premene, prim. *korenj*. Tudi večina besed na *-lje* ima v rod. mn. premeno *-lij: nasélje – nasélij*. – Pri besedah *dnò* in *tlà* se v rod. mn. in dv. vriva *a: dán* (poleg običajnejšega *dnòv/dnôv*) in *tál*. Prim. str. 58.

Prednost celovite formalizirane obravnave oblikoslovja je torej urejenost (četudi na prvi pogled razdrobljenih) rezultatov, na osnovi katere je mogoče tudi slovnični opis pripraviti urejeno, ločeno po prepoznanih značilnostih in podprto s sodobnim gradivom. Že uporaba v sodobnem jeziku pogostih zgledov olajša razumevanje obravnavanih slovničnih pojavov. Tako bi preprosteje

zapisali,[21] da se samostalniki srednjega spola na *-o* pregibajo drugače, če pred o-jem stoji kombinacija nezvočnika in zvočnika *v, l, n, r* ali *m;* npr. samostalniki *ministrstvo, geslo, okno, jutro, pismo*. Pri teh samostalnikih se v rodilniku dvojine in množine med nezvočnik in zvočnik vrine *-e-* (npr. *geslo – gesel* in ne *geslo – gesl,* za razliko od običajnega pregibanja *delo – del, mesto – mest*). Posebnosti pregibanja samostalnikov, ki se končajo na *-je*, lahko nato opis obravnava ločeno, ker se v resnici obnašajo precej drugače. Prav tako se lahko ločeno obravnavajo specifike naglaševanja.

V primeru, da bodo novi slovnični opisi zasnovani za digitalno obliko, kar je zaželeno oz. pričakovano, je mogoče v besedilo dodati povezave na celotne paradigme za obravnavane zglede, kakor tudi na sezname vseh samostalnikov, ki se pregibajo na v razdelku opisani način. Podatkovna povezljivost na eni in kvantiteta na drugi strani lahko pomembno olajšata razumevanje slovničnih pojavov, sploh za jezikovnodidaktične namene. V smislu povezljivosti je kot zadnjo prednost predstavljene strojne obravnave mogoče izpostaviti sopostavitev lastnoimenskih samostalnikov ob občnoimenske. Vprašanja pregibanja lastnih imen se v Slovenski slovnici sicer pojavljajo, vendar ne sistematično, in velik del vprašanj, zlasti o pregibanju tujih lastnih imen, se prepušča v obravnavo pravopisnim priročnikom. Z vidika uporabnika je zaželeno, da slovnični opis pokrije tudi lastnoimenske podatke, splošnejšo slovnično obravnavo pa je mogoče povezati s problemskim pristopom, kot ga predvideva že omenjeni koncept Slogovnega priročnika (Krek in dr. 2013).

6 SKLEP IN NADALJNJE DELO

Relativno preprost pristop k strojni obravnavi oblikoslovnih vzorcev, kot so ga napovedali (Dobrovoljc in dr. 2015b) v sklopu priprav na novi slovar sodobne

[21] Nikakor ni namen članka ponuditi alternativni slovnični opis za katero koli od prepoznanih jezikovnih značilnosti, za slednje je potreben celovitejši razmislek in več jezikoslovnih analiz gradiva kot izhodišče oblikovanim trditvam. Na tem mestu želiva le ponazoriti, kako lahko izgradivna urejenost rezultatov pripomore k urejenosti opisa.

slovenščine, se po prvi implementaciji izkazuje za plodnega in učinkovitega, pod pogojem, da mu sledi ročna jezikoslovna analiza, podprta z razumevanjem trenutnega ustroja leksikona Sloleks. V nadaljevanju dela bodo samostalniški vzorci prerazvrščeni in leksikonsko pridobljeni podatki posodobljeni. Z nadgrajenim programom bodo izluščene, nato pa ročno urejene paradigme za ostale besedne vrste. Sledilo bo dopolnjevanje gradiva s podatki iz korpusa oz. korpusov; v prvem koraku je v načrtu uporaba korpusa Gigafida 2.0 (Krek in dr. 2016), postopek je seveda mogoče uporabiti tudi na drugih korpusih, pri čemer bo posebna pozornost namenjena težavam na ravni nestandardnih in redkih arhaičnih oz. stilnih variant. Predvideno je, da bodo za ta del potrebne dodatne jezikoslovne analize, ki bodo opredelile metodologijo luščenja in de facto pojavnost redkih oblik. Po dopolnitvi s korpusnimi podatki bo pripravljen končni nabor vzorcev, ki ga bomo vpisali v leksikonsko bazo, strokovni javnosti pa bo dokumentirani in strukturirani seznam na voljo tudi na repozitoriju Clarin.si. Zadnji korak je dopolnitev leksikonskega vmesnika, ki bo omogočil, da s klikom na izpisano kodo vzorca uporabnik dostopa do zbranega nabora vseh ustrezajočih lem. Od tam je mogoče vzpostaviti tudi povezave na vire, ki lahko določene oblikoskladenjske specifike natančneje obravnavajo.

ZAHVALA

Predstavljeno znanstvenoraziskovalno delo je rezultat projekta ‘Nova slovnica sodobne standardne slovenščine: viri in metode’ (šifra ARRS: J6-8256), ki ga sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

LITERATURA

Arhar, Š. (2009): Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54 (3–4): 43–56.

Arhar, Š. in Holozan, P. (2009): Leksikalna podatkovna zbirka ASES (Amebisov skupni elektronski slovar). V V. Mikolič (ur.): *Jezikovni korpusi v medkulturni komunikaciji*: 30–51. Koper: Univerza na

Primorskem, Znanstveno-raziskovalno središče, Založba Annales: Zgodovinsko društvo za južno Primorsko.

Arhar Holdt, Š., Dobrovoljc, K. in Popič, D. (2013): Reprezentacija standardnega in nestandardnega v virih SSJ. V A. Žele (ur.): *Družbena funkcijskost jezika (vidiki, merila, opredelitve)*: 19–27. Ljubljana: Znanstvena založba Filozofske fakultete.

Čibej, J., Arhar Holdt, Š., Erjavec, T. in Fišer, D. (2016): Razvoj učne množice za izboljšano označevanje spletnih besedil. V T. Erjavec in D. Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika:* 40–46. Ljubljana: Znanstvena založba Filozofske fakultete.

Dobrovoljc, K. (2015): Oblikoslovne informacije v sodobnih slovarskih priročnikih. V V. Gorjanc in dr. (ur.): *Slovar sodobne slovenščine: problemi in rešitve:* 64–79. Ljubljana: Znanstvena založba Filozofske fakultete.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T. in Romih, M. (2015a): *Morphological lexicon Sloleks 1.2*. Ljubljana: Slovenian Language Resource Repository CLARIN.SI, 2015. http://hdl.handle.net/11356/1039

Dobrovoljc, K., Krek, S. in Erjavec, T. (2015b): Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc in dr. (ur.): *Slovar sodobne slovenščine: problemi in rešitve:* 80–105. Ljubljana: Znanstvena založba Filozofske fakultete.

Erjavec, T. in Krek, S. (2008): Oblikoskladenjske specifikacije in označeni korpusi JOS. V T. Erjavec in J. Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije: zbornik 11. mednarodne multikonference Informacijska družba - IS 2008*: 49–53. Ljubljana: Institut Jožef Stefan.

Erjavec, T., Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S. in Velušček, A. (2008): *Specifikacije za leksikon besednih oblik – projekt*

Sporazumevanje v slovenskem jeziku, kazalnik 3. Kamnik. Dostopno prek: http://projekt.slovenscina.eu/Vsebine/Sl/Kazalniki/K3.aspx (2. 12. 2018).

Erjavec, T. (2015): The IMP historical Slovene language resources. *Language resources and evaluation*, 49 (3): 753–775.

Fišer, D., Erjavec, T. in Ljubešić, N. (2016): JANES v0.4: korpus slovenskih spletnih uporabniških vsebin. V: D. Fišer (ur.). *Računalniško posredovana komunikacija*, *Slovenščina 2.0*, 4 (2): 67–994. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S., ur. (2015): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete.

Gorjanc, V. (2017): *Nije rečnik za seljaka*. Beograd: Biblioteka XX vek.

Grčar, M., Krek, S. in Dobrovoljc, K. (2012): Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije: zbornik 15. mednarodne multikonference Informacijska družba:* 89–94. Ljubljana: Institut Jožef Stefan.

Krek, S., Dobrovoljc, H., Dobrovoljc K. in Popič, D. (2013): Online style guide for Slovene as a language resources hub. V I. Kosem in dr. (ur.): *Electronic lexicography in the 21st century: thinking outside the paper*. *Proceedings of eLex 2013 Conference:* 379–391. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut.

Krek, S., Gantar, P., Arhar Holdt, Š. in Gorjanc, V. (2016): Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. V T. Erjavec in D. Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika:* 200–202. Ljubljana: Znanstvena založba Filozofske fakultete.

Ljubešić, N., Erjavec, T. in Fišer, D. (2014): Standardizing tweets with character-level machine translation. V A. Gelbukh (ur.): *Computational linguistics and intelligent text processing:* 164–175. Heidelberg [etc.]: Springer.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S, in Škrjanec, I. (2015): Predicting the level of text standardness in user-generated content. *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015*: 371–378. Hissar, Bulgaria.

Ljubešić, N. in Erjavec, T. (2016): Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. *Proceedings of Language Resources and Evaluation Conference (LREC) 2016*: 1527–1531. Portorož, Slovenia.

Logar, N., Grčar, M., Brakuš, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Ljubljana: Fakulteta za družbene vede.

Pollak, S. in Božinovski, B. (2014): Luščenje borzne terminologije. V T. Erjavec in J. Žganec Gros (ur.): *Jezikovne tehnologije: zbornik 17. mednarodne multikonference Informacijska družba*: 114–119. Ljubljana: Institut Jožef Stefan.

Rejc, R. (2017): Generiranje slovenskih besednih oblik s pomočjo strojnega učenja [diplonsko delo]. Dostopno prek: https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=91151 (2. 12. 2018).

Scherrer, Y. in Erjavec, T. (2016): Modernising historical Slovene words. *Natural language engineering*, 22 (6): 881–905.

Slovar slovenskega knjižnega jezika (2., dopolnjena in deloma prenovljena izd., elektronska objava, 2014). Ljubljana: SAZU in Inštitut za slovenski

jezik Frana Ramovša ZRC SAZU. Dostopno prek: www.fran.si (oktober 2018).

Stritar, M. in Dobrovoljc, K. (2013): Korpusi na poti v šole: jezikovnotehnološko izpopolnjevanje učiteljev. *Slovenščina 2.0*, 1 (1): 181–194.

Toporišič, J. (2004): *Slovenska slovnica*. Maribor: Obzorja.

Vintar, Š. (2015): Terminologija v spletnih forumih. V D. Fišer (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih:* 69–74. Ljubljana: Znanstvena založba Filozofske fakultete.

MORPHOLOGICAL PATTERNS IN THE SLOLEKS LEXICON OF SLOVENE: AN INITIAL SET OF PATTERNS FOR NOUNS

The paper presents the first step to expanding the Sloleks lexicon of Slovene with morphological patterns, starting with nouns. In the first phase, the patterns were extracted automatically from the lexicon based on a selection of differentiating characteristics (morphosyntactic tags and variable word parts). This was followed by a manual categorization during which we (a) separated patterns that are either systemic or based on actual language use from examples extracted because of noise attributable to either the extraction method or inconsistencies in Sloleks; (b) arranged patterns into groups based on their content and relatedness; (c) analyzed and more clearly defined form variability, with both standard and non-standard word forms; (d) propose future steps for the further development of the extraction method and lexicon upgrades. The result is a set of formalized morphological patterns for (common and proper) nouns containing 10 groups (64 patterns) for masculine nouns, 9 groups (29 patterns) for feminine nouns and 8 groups (20 patterns) for neuter nouns. The preparation of the set of formalized patterns also resulted in numerous suggestions on how to upgrade the lexicon, while a machine-focused view of morphological flection offers opportunities to improve the current grammatical description of Slovene. As part of our future work, we intend to expand the set of patterns with other parts of speech and corpus-based material. The final categorization of patterns will be included in the Sloleks lexicon, and the patterns will also be published on the CLARIN.SI repository in a machine-readable format.

Keywords: Sloleks, word form lexicon, morphological patterns, noun, Slovene