Strojno luščenje medbesednih povezav v oblikoslovnem leksikonu Sloleks 2.0

Jaka ČIBEJ
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
jaka.cibej@ff.uni-lj.si

Abstract

In the paper, we present an automatic rule-based approach to extracting word relations between morphologically related Slovene words (e.g. *hraber* 'brave' – *hrabrost* 'bravery') in order to expand the number of word relations included in Sloleks 2.0, the Slovene Morphological Lexicon. The approach relies on a set of rules designed bottom-up using predictable word parts that are used in Slovene word formation. The method resulted in approximately 66,000 extracted word relations, and preliminary evaluations show that between 75 and 80 % are adequate, with certain rules being more reliable. We provide an overview of the most productive and most problematic rules and describe our plans for future work in the conclusion.

Ključne besede: medbesedne povezave, povezovalna pravila, besedni deli, besedotvorje, računalniško jezikoslovje

Keywords: word relations, word relation rules, word parts, word formation, computational linguistics

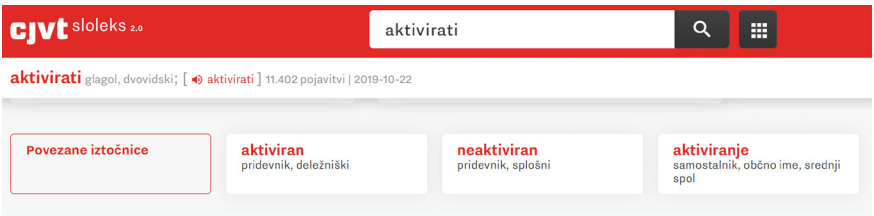
1 Uvod

Slovenski oblikoslovni leksikon Sloleks 2.0[1] je trenutno najobsežnejša odprto dostopna baza s podatki o slovenskih besednih oblikah in njihovih oblikoskladenjskih značilnostih. V različici 2.0 vsebuje 100.802 iztočnici in 2.792.003 besedne oblike, vsaki pa je pripisana tudi oblikoskladenjska oznaka po sistemu MULTEXT-East v6,[2] ki nakazuje besedno vrsto (npr. samostalnik), druge slovnične značilnosti oblike (za samostalnike npr. občno- ali lastnoimenskost, spol, število, sklon, živost) in frekvenčne podatke iz korpusa pisne standardne slovenščine Gigafida 2.0 (Krek et al. 2020).

Poleg podatkov o sami iztočnici oz. obliki vsebuje tudi podatke o povezanih iztočnicah – iztočnica ima lahko navedene povezave z drugimi besedotvorno povezanimi besedami (npr. *pisati* → *pisanje*), a je število povezav v trenutni različici nekoliko omejeno: Dobrovoljc et al. (2015) navajajo, da različica Sloleksa 1.2 (ki je po naboru iztočnic enaka različici 2.0) z vidika besedotvornih povezav vsebuje le nekatere recipročne povezave, npr. med samostalnikom in izpeljanim svojilnim pridevnikom (*kruh* → *kruhov*), med glagolom in izpeljanim glagolnikom (*briti* → *britje*), med pridevnikom in izpeljanim samostalnikom na *-ost* (*zarjavel* → *zarjavelost*), med glagolom in izpeljanim deležjem (*začeti* → *začenši*), med glagolom in izpeljanim deležnikom (*ujeti* → *ujet*), med pridevnikom in izpeljanim prislovom (*navihan* → *navihano*), med pridevnikom in izpeljanim elativom (*lep* → *prelep*), med prislovom in izpeljanim elativom (*glasno* → *preglasno*) ter med lemo in njeno okrajšavo (*gospodična* → *gdč.*). V trenutnem vmesniku za Sloleks, ki je dostopen od leta 2019, lahko uporabnik prehaja z oblik izbrane iztočnice na oblike povezanih iztočnic s pomočjo ploščic (Slika 1), opaziti pa je mogoče nekatere nedoslednosti pri navajanju povezanih iztočnic: obstajata npr. povezavi *aktiviran* → *aktivirati* ter *aktivirati* → *neaktiviran*, ni pa povezave *aktiviran* → *neaktiviran*. Prostora za izboljšave je torej še veliko.

1 Spletni vmesnik Slovenskega oblikoslovnega leksikona Sloleks 2.0: https://viri.cjvt.si/sloleks/slv/.

2 MULTEXT-East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.



Slika 1: Povezane iztočnice za iztočnico *aktivirati* v Sloleksu 2.0.

Povezave v leksikonu bi bilo smiselno dopolniti iz več razlogov: kot prvo, celosten nabor povezanih iztočnic v oblikoslovnem leksikonu je lahko zelo koristen pri gradnji derivacijskih morfoloških mrež in jezikovnih (slovarskih) virov, ki lahko iz baze leksikona črpajo povezave oz. predloge za povezane iztočnice, in pri pripravi učnih gradiv za usvajanje besedišča pri učenju slovenščine kot drugega/tujega jezika. Podatki lahko koristijo tudi razvoju jezikovnih tehnologij za slovenščino, npr. za sisteme za razreševanje anafor, za krnilnike za slovenščino in za morebitne distribucijskosemantične modele za rudarjenje podatkov ali indeksiranje dokumentov.

Ročno dopolnjevanje povezav v leksikonu je časovno zelo potratno, zato je dobro imeti vzpostavljen formaliziran sistem, ki strojno povezuje besedotvorno sorodne enote. Dobro je, da je sistem zasnovan dovolj robustno, da je uporaben tudi pri morebitnih drugih nalogah, npr. za širjenje oblikoslovnega leksikona z novimi iztočnicami iz korpusov, strojno tvorjeni kandidati pa lahko služijo tudi kot vir predlogov za poimenovalne kandidate za nove pojavnosti, kar je koristno npr. za prevajalce, terminologe in pisce besedil.

V okviru obdelave naravnega jezika za druge jezike že obstajajo raziskave na temo strojnega generiranja morfoloških derivacijskih mrež z različnimi pristopi (na podlagi strojnega učenja, pravil ali hibridnih modelov): npr. Lango et al. (2020) za poljščino in španščino, Zeller et al. (2013) za nemščino, Lignos et al. (2009) za angleščino in nemščino ter Ševčíková (2018) za češčino. Za slovenščino je bilo izvedenih že precej podrobnih jezikoslovnih raziskav in teoretičnih obravnav besedotvorja: poleg Slovenske slovnice (Toporišič 2004), ki besedotvorju namenja ločeno poglavje in navaja nabor predpon in

pripon ter nudi splošno razlago besedotvornih postopkov v slovenščini (izpeljava, sestavljanje, zlaganje, sklapljanje), je treba omeniti še Vidovič Muha (1988), ki nudi bolj celosten pregled besedotvornih postopkov v slovenščini na primeru zloženk in izpeljank iz njih (npr. *častihlepen*, *častihlepnež*, *častihlepnik*, *častihlepnost*). Jakopin et al. (2009) analizirajo besedne dele v novejši slovenski leksiki (tudi iz spletnih besedil), v zadnjem času pa je besedotvorna problematika v slovenskem jezikoslovju obravnavana skozi lečo stopenjskega besedotvorja (Kern 2017), ki obravnava skupine tvorjenk, razporejene ob netvorjeni besedi (npr. *avantgarda*, *avantgardist*, *avantgardističen*, *avantgardističnost*); tovrstno razvrščanje tvorjenk po stopnjah je značilno za slovenščino in za nekatere druge slovanske jezike (za poljščino npr. glej Skarżyński 2000). Kern (2010) stopenjsko besedotvorje opredeli kot del besedotvorja, katerega namen je izdelati pregleden nabor tvorjenk glede na netvorjeno besedo skupaj z analizo, v kolikšnim meri so korenske besedotvorne podstave besedotvorno produktivne, katere besedne vrste tvorijo ipd. Stopnje tvorjenosti je med drugim mogoče predstaviti v t. i. tvorbenem modelu: npr. Kern (2011, 2017) verigo stopenj *stopiti – odstopiti – odstop* predstavi z modelom V,V,S (glagol, glagol, samostalnik). Še dodatno je mogoče tvorbeni model opredeliti z nizom obrazil, ki se v tvorbenem modelu uporabijo (npr. 'X- + -en + (ne-) + -ost' za *neopaznost*, glej Kern 2020: 74), iz različnih tvorbenih modelov pa je mogoče tvoriti besedne družine (*šikana – šikanozen*/*šikanirati – šikaniranje*, glej npr. Stramljič Breznik 2020: 80). Kombinatorika morfemskih obrazil (oz. morfotaktika) je v tem pogledu tudi v slovenskem jezikoslovju še podraziskana, trenutno pa prav tako še ni raziskav, ki bi problematiko besedotvorja obravnavale jezikovnotehnološko, zato (odprto dostopna) baza s podatki o besedotvornih pravilih v strojno berljivi obliki še ne obstaja. Pričujoča raziskava ima torej dva poglavitna cilja: (a) nabor medbesednih povezav, s katerim bo mogoče obogatiti Slovenski oblikoslovni leksikon Sloleks, in (b) prvi korak k formalizaciji besedotvornih podatkov o slovenščini v strojno berljivi obliki.

V prispevku najprej predstavimo metodologijo izdelave povezovalnih pravil na podlagi besednih delov (razdelek 2) ter luščilni

algoritem (razdelek 3), nato pa predstavimo nabor približno 66.000 povezav in opravimo preliminarno evalvacijo luščilne točnosti izdelanih pravil (razdelek 4). V zaključku (razdelek 5) strnemo ugotovitve in začrtamo smernice za prihodnje delo.

2 Metodologija

Razvoj algoritma za strojno luščenje povezav med leksikonskimi enotami je potekal v več korakih. V prvem koraku smo pripravili nabor besednih delov, na podlagi katerih smo v drugem koraku izdelali nabor povezovalnih pravil. V zadnjem koraku smo pravila uporabili za luščenje in nazadnje opravili evlavacijo njihove uspešnosti na podlagi stratificiranega vzorca izluščenih medbesednih povezav. Vsi koraki so podrobneje opisani v nadaljevanju.

2.1 Priprava nabora besednih delov

Za izhodišče smo pregledali vse pripone in predpone, ki so navedene v poglavju Besedotvorje v Slovenski slovnici (Toporišič 2004: 143–232). Razvezali smo dvojnice in variante ter odstranili morebitne naglase in ločila (npr. *-(á)lec → alec*, *lec*; *-inja/-ínja → inja*) ter zabeležili, pri kateri besedni vrsti se pojavljajo. V prispevku v nadaljevanju, ko opisujemo strojno luščenje, govorimo o besednih delih, saj jih obravnavamo formalizirano (ne ločujemo jih npr. glede na pomen) in zgolj na podlagi površinskih oblik, zato se naše delitve besed ne prekrivajo nujno z delitvami, kot so pojmovane v slovenskih besedotvornih raziskavah. Ko navajamo predpone in pripone, kot so navedene v Slovenski slovnici, jih navajamo z vezajem (-). Besedne dele, kot smo jih uporabili pri luščenju, pa navajamo s podčrtajem (_).

Nato smo izvedli dve luščenji lem iz Sloleksa 2.0: v prvem smo izluščili in razcepili vse leme, ki se začnejo s katerimkoli besednim delom iz nabora začetnih besednih delov (npr. *pri_*, *pre_*, *od_*, *nad_*), v drugem pa vse, ki se končajo s katerimkoli besednim delom iz nabora končnih besednih delov (npr. *_išče*, *_anje*, *_ik*). V primerih, ko je bilo besedo mogoče razcepiti na več načinov, smo upoštevali najdaljši možni besedni del (lemo *provokator* smo npr. razcepili

kot *provok_ator*, ne *provokat_or*; podobno tudi *pred_staviti* namesto *pre_dstaviti*). Na ta način smo zmanjšali delež napačnih cepljenj (v nasprotnem primeru bi lahko npr. vse besede s končnim besednim delom *_anje* pristale pod končnim besednim delom *_je*, kar bi bilo za analizo kontraproduktivno).

V drugem luščenju smo poskušali zajeti tudi morebitne besedne dele, ki niso zabeleženi v Slovenski slovnici. Leme smo cepili na začetne in končne dvo-, tri- in štiričrkovne besedne dele. Opravili smo pregled tako dobljenih razdeljenih enot in besedne dele bodisi potrdili kot relevantne (tj. ali se kot relevantni besedni deli pojavljajo v iztočnicah Sloleksa 2.0) ali pa smo jim pripisali, da enot s tovrstnim besednim delom v leksikonu nismo našli. Končni nabor je znašal 1.013 besednih delov[3] (Tabela 1), od tega 359 končnih in 654 začetnih.

Tabela 1: Število besednih delov, uporabljenih za pisanje povezovalnih pravil.

Vrsta besednega dela	Vse besedne vrste	Samostalniki	Samostalniki moškega spola	Samostalniki ženskega spola	Samostalniki srednjega spola	Pridevniki	Glagoli	Prislovi
Končni besedni deli	140	-	57	9	9	31	7	27
Sestavljeni končni besedni deli	219	-	52	71	25	50	16	5
Začetni besedni deli	367	93	-	-	-	93	90	91
Sestavljeni začetni besedni deli	287	47	-	-	-	78	129	33

Na tej točki je treba omeniti, da smo nekatere besedne dele, ki so bili v Slovenski slovnici navedeni kot samostojni (npr. *-ovati*, *-janski*), razdelili in jih kategorizirali kot sestavljene besedne dele (npr. *_ov_ati*, *_j_an_ski*). To je še posebno pomembno v primerih, ko gre za delno prekrivnost z drugimi besednimi deli (*_ov_ati – _ati*; *_an_ski – _ski*). Na ta način smo lahko dosegli delitev besed, ki je bolj konsistentna

3 Prve dele zloženk (npr. *geo_politika*) smo med pregledom izluščenih iztočnic sicer beležili, a jih pri pisanju povezovalnih pravil v tej različici luščilnega algoritma še nismo upoštevali, saj zahtevajo drugačno obravnavo in temeljitejšo analizo.

med različnimi besednimi vrstami: namesto delitev *ion-izirati* in *ioniz--acija*, ki bi bili rezultat obravnave s priponami in predponami po Slovenski slovnici, smo tako dobili delitvi *ion_iz_ir_ati* in *ion_iz_ac_ij_a*, ki imata v tem primeru enak osrednji del (*ion*).

Tovrstna delitev omogoča tudi manjši nabor povezovalnih pravil, saj je formaliziran pristop bolj ekonomičen in dopušča, da eno samo pravilo uporabimo za več različnih kombinacij besednih delov: za povezovanje besed *ion_iz_ir_ati* in *ion_iz_ac_ij_a* lahko npr. uporabimo enako pravilo kot za par *oper_ir_ati* in *oper_ac_ij_a* ne glede na to, da se kombinacija končnih besednih delov, ki sledijo osrednjemu delu, med paroma nekoliko razlikuje (*_iz_ir_ati – _ir_ati*, *_iz_ac_ij_a – _ac_ij_a*).

Dodali smo tudi različice, ki so bile v Slovenski slovnici le implicitne oz. niso bile navedene, ker so obravnavane kot podaljšava osnove: *_j_ev_ski* (*hipi_j_ev_ski*) npr. ni bil eksplicitno naveden, a je bil impliciran pod *-evski,* podobno tudi *_j_ev* (*urar_j_ev*), ki je impliciran z *-ev*). S tem smo dosegli še večjo stopnjo formaliziranosti, ki je nujno potrebna za pisanje pravil in njihovo luščilno točnost.

Visoko število začetnih besednih delov pri glagolih je treba pripisati dejstvu, da smo za razliko od Slovenske slovnice, ki navaja samo posamezne predpone, pri glagolih upoštevali tudi kombinacije, v katerih se lahko pojavljajo začetni besedni deli (npr. *raz_po_red_iti*, *pred_po_stav_iti, po_raz_del_iti*). Na te kombinacije opozori npr. Jakopin (1971: 1–2): »[...] skoraj vsi osnovni glagoli se združujejo z domala vsemi produktivnimi predponami, nekateri pa tudi z dvema hkrati (npr. s-pre-hoditi)«, omenja pa jih tudi Kern (2011: 130) v analizi besedotvornih sklopov glagola *stopiti* (npr. *pred_v_stop_en*). Te kombinacije smo prav tako pridobili s pregledom izluščenih iztočnic iz Sloleksa 2.0, kategorizirali pa smo jih kot sestavljene začetne besedne dele. Enake kombinacije se seveda lahko pojavljajo tudi pri drugih besednih vrstah (npr. *po_raz_del_it_ev*), a ker eden od luščilnih algoritmov (glej razdelek 2.3.1) kot izhodišče za luščenje povezav vzame glagol in njegovo delitev nato prenese tudi na ostale povezane besedne vrste (npr. *po_raz_del_it_v_en*), kombinacij v naboru nismo navajali pri vseh besednih vrstah.

Skupno 67 besednih delov (33 končnih besednih delov za moške samostalnike, 16 končnih besednih delov za ženske samostalnike, 8 končnih besednih delov za samostalnike srednjega spola, 4 končne besedne dele za glagole in 5 končnih besednih delov za prislove) ni bilo vključenih v nabor za pisanje pravil, in sicer iz več razlogov: (a) ker zanje niti v Sloleksu 2.0 niti v korpusu Gigafida 2.0 nismo našli primerov (npr. *-ataj* za samostalnike moškega spola: *vozatáj*; *-kljat* za pridevnike: *rumenkljat*; *-leti* za glagole: *frleti*), (b) ker je bilo primerov malo (2 ali manj) in luščenje s pravilom ne bi bilo produktivno (npr. *-cat* pri *sam-cat*, *prav-cat*), in (c) ker se je besedni del nanašal le na imenske entitete, ki jih pri luščenju trenutno nismo upoštevali (npr. *-j* za pridevnike: *Slovenj*).

2.2 Povezovalna pravila za morfološko povezane besede

Ko smo določili končni nabor besednih delov, smo z njimi izvedli še tretje luščenje iz Sloleksa 2.0 in pridobili kandidate, ki so razcepljeni glede na končni nabor besednih delov. Nato smo ročno pregledali kandidate v vsaki izluščeni skupini in izdelali pravila za medbesedne povezave, ki smo jih pozneje uredili v hierarhijo glede na besedno vrsto izvorne in povezane besede ter glede na besedne dele, na podlagi katerih pravilo deluje. Primere pravil prikazuje Tabela 2.

Identifikacijska koda pravila je sestavljena iz besedne vrste oz. oblikoskladenjskih značilnosti izvorne in povezane besede[4] po označevalnem sistemu MULTEXT-East v6 (https://nl.ijs.si/ME/V6/msd/html/msd-sl.html) ter identifikacijskih številk, ki ponazarjajo skupine in podskupine pravil v sklopu celotne hierarhije. Samo pravilo nakazuje, da vzamemo iztočnico določene besedne vrste z določenim končnim besednim delom (npr. *[G]_ati*, glagol s končnim besednim delom *_ati*). Če odstranimo končni del izvorne besede in ga nadomestimo z drugim končnim delom (npr. *[G]_anj_e*, preostanek glagola in končni besedni del *_anj_e*), dobimo povezano iztočnico s ciljno besedno vrsto. Tabela 3 prikazuje vsa pravila za določanje povezav med glagoli

4 Pri samostalnikih sta poleg besedne vrste navedena še občnoimenskost/lastnoimenskost in spol.

Tabela 2: Primeri povezovalnih pravil za luščenje medbesednih povezav.

Identifikacijska koda	Pravilo	Besedna vrsta izvorne besede	Besedna vrsta povezane besede	Primer
Som.Som.1.1	[S] → [S]_ec	Som	Som	dvor → dvorec
Som.Som.1.2.1	[S] → [S]_av_ec	Som	Som	list → listavec
G.Sos.3.1	[G]_ati → [G]_anj_e	G	Sos	pisati → pisanje
G.Sos.3.2.1	[G]_eti → [G]_enj_e	G	Sos	goreti → gorenje
Soz.P.1.1	[S]_a → [S]_ski	Soz	P	absorpcija → absorpcijski
Soz.P.3.5	[S]_a → [S]_ar_en	Soz	P	disciplina → disciplinaren
P.Soz.7.1.1	[P] → [P]_ost	P	Soz	dokazan → dokazanost
P.R.1	[P]_en → [P]_n_o	P	R	normalen → normalno

('G') in pridevniki ('P') iz skupine 1 (pridevniki na *_oč*/*_eč*). Ta skupina se deli na podskupini 1 (pridevniki na *_oč*) in 2 (pridevniki na *_eč*), ki vsebujeta posamezna povezovalna pravila (npr. G.P.1.1.1 za povezavo med glagoli na *_ati* in pridevniki na *_oč*, npr. *smej_ati* → *smej_oč*).

Tabela 3: Prva skupina pravil za povezovanje glagolov in pridevnikov.

Identifikacijska koda	Pravilo	Besedna vrsta izvorne besede	Besedna vrsta povezane besede
G.P.1.1.1	[G]_ati → [G]_oč	G	P
G.P.1.1.2	[G]_eti → [G]_oč	G	P
G.P.1.1.3	[G]_sti → [G]_oč	G	P
G.P.1.1.4	[G]_ati → [G]_aj_oč	G	P
G.P.1.1.5	[G]_iti → [G]_uj_oč	G	P
G.P.1.1.6	[G]_eti → [G]_uj_oč	G	P
G.P.1.1.7	[G]_ev_ati → [G]_uj_oč	G	P
G.P.1.1.8	[G]_ov_ati → [G]_uj_oč	G	P
G.P.1.1.9	[G]_iti → [G]_ij_oč	G	P
G.P.1.2.1	[G]_eti → [G]_eč	G	P
G.P.1.2.2	[G]_iti → [G]_eč	G	P
G.P.1.2.3	[G]_ati → [G]_eč	G	P

Tabela 4: Število pravil v posameznih skupinah v prvi različici nabora pravil.

Skupina pravil	Vrsta medbesedne povezave	Število pravil
G.P	Povezava med glagolom in pridevnikom	58
G.Sos	Povezava med glagolom in občnim samostalnikom srednjega spola	29
G.Soz	Povezava med glagolom in občnim samostalnikom ženskega spola	51
G.Som	Povezava med glagolom in občnim samostalnikom moškega spola	62
G.R	Povezava med glagolom in prislovom	19
P.P	Povezava med dvema pridevnikoma	7
Som.P	Povezava med občnim samostalnikom moškega spola in pridevnikom	78
Soz.P	Povezava med občnim samostalnikom ženskega spola in pridevnikom	57
Sos.P	Povezava med občnim samostalnikom srednjega spola in pridevnikom	14
Som. Som	Povezava med dvema občnima samostalnikoma moškega spola	41
Soz.Som	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom moškega spola	33
Sos.Som	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom moškega spola	2
Soz.Sos	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom srednjega spola	11
Som.Sos	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom srednjega spola	12
Som.Soz	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom ženskega spola	29
Soz.Soz	Povezava med dvema občnima samostalnikoma ženskega spola	18
Sos.Soz	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom ženskega spola	5
P.Som	Povezava med pridevnikom in občnim samostalnikom moškega spola	18
P.Soz	Povezava med pridevnikom in občnim samostalnikom ženskega spola	29
P.Sos	Povezava med pridevnikom in občnim samostalnikom	3
P.R	Povezava med pridevnikom in prislovom	3
Sos.Sos	Povezava med dvema občnima samostalnikoma srednjega spola	4

Vseh pravil, ki so bila uporabljena za luščenje povezav, je v prvi različici nabora 583 (Tabela 4). Med pregledom izluščenih enot smo

zabeležili tudi nekaj pravil, ki vključujejo lastnoimenske samostalnike (‘Slm’/‘Slz’/‘Sls’), a jih v nabor še nismo vključili, saj potrebujejo ločeno in natančnejšo obravnavo, zlasti v primeru tujih lastnih imen (*Shakespeare* → *Shakespearov*). Prav tako v trenutno različico še nismo vključevali povezav, ki vsebujejo druge besedne vrste po sistemu MULTEXT-East v6, npr. številke (‘K’) in zaimke (‘Z’).

Hierarhija torej ni izčrpna in je zasnovana tako, da je vanjo mogoče dodajati nova pravila oz. urejati in prerazporejati obstoječa. Omeniti je treba tudi, da so povezave lahko recipročne in ne upoštevajo nujno smeri besedotvornega postopka, kot je določena v jezikoslovnih raziskavah; iz glagola *predsednikovati* npr. lahko pridobimo povezano iztočnico *predsednik*. V določenih primerih je povezava do iste ciljne besede (vsaj v trenutni različici luščilnega algoritma) lahko ustvarjena po več različnih pravilih (npr. *liofil_iz_ir_ati* → *liofil_iz_ir_anj_e* kot povezava med glagolom in samostalnikom, *liofil_iz_ir_an* → *liofil_iz_ir_an_je* kot povezava med pridevnikom in samostalnikom.

2.3 Algoritem vzpostavljanja medbesednih povezav

Na podlagi nabora pravil smo izdelali algoritem, ki kot izhodišče vzame iztočnice iz Sloleksa 2.0 skupaj z njihovimi oblikoskladenjskimi značilnostmi, na podlagi pravil pa iz njih tvori ciljne iztočnice in jih preveri v leksikonu. Če je tako nastala iztočnica prisotna v leksikonu, algoritem medbesedno povezavo izpiše kot veljavno. V nasprotnem primeru morebitno ciljno iztočnico zabeleži kot nenajdeno. Na ta način algoritem pridobi nabor povezav med izvorno in ciljno besedo, a ker povezave lušči hierarhično, je mogoče tako pridobljene povezave razvrstiti tudi v verige oz. drevesa po vzoru stopenjskega besedotvorja (Kern 2010). V tem prispevku se osredotočamo samo na izluščene povezave, ne pa na njihova medsebojna razmerja.

Algoritem je medbesedne povezave izvažal nekoliko drugače glede na izhodišče, ki je vključevalo bodisi glagole (razdelek 2.3.1) bodisi druge besedne vrste (razdelek 2.3.2). Oba postopka podrobneje predstavljamo v nadaljevanju.

2.3.1 Luščenje z izhodiščem pri glagolih

Povezave z glagoli smo obravnavali ločeno, saj je njihova delitev na besedne dele nekoliko bolj predvidljiva in obenem zelo regularna, poleg tega pa je glagolov v Sloleksu 2.0 le okrog 10.000, kar je še obvladljivo za ročni pregled. V prvem koraku smo avtomatsko razcepili vse glagolske iztočnice na morebitne začetne (*na_*), osrednje (*_pis_*) in končne dele (*_ati*), nato pa smo jih ročno pregledali ter popravili morebitne napačne delitve in tako pridobili nabor 2.621 potencialnih osrednjih delov.

V naslednjem koraku smo iz osrednjih delov s pomočjo nabora začetnih delov (oz. kombinacij začetnih delov) in končnih delov (kot smo jih našteli v Tabeli 1) tvorili glagolske kandidate in vsakega najprej preverili v leksikonu – če je bil kandidat med iztočnicami, je algoritem iz glagola glede na nabor pravil ustvaril nove besede, jih znova preveril v leksikonu in na ta način potrdil povezavo med glagolom in ciljno besedo. Za vsako ciljno besedo, ki jo je algoritem potrdil, je iz nje rekurzivno znova ustvaril nove besede na podlagi istega nabora pravil in ponavljal postopek, dokler ni izčrpal vseh možnosti, nato pa se je vračal k prejšnjim besedam in tvoril nove kandidate. Izsek, ki ponazarja delovanje algoritma za luščenje medbesednih povezav z glagolskim izhodiščem, je prikazan na Sliki 2 – v drevesni strukturi so izpisani glagoli in povezane iztočnice skupaj s pravili, po katerih je bila povezava vzpostavljena. Zaradi konciznosti so izpisane le nekatere izluščene povezave (izpuščene povezave so označene z […], zvezdica (*) pa označuje kandidate, ki niso vključeni v leksikon).

Algoritem v zgornjem primeru začne z osrednjim delom *_pis_*, ki mu nato pripenja različne končne (*_ati*, *_ov_ati, _eti, _iti*) in začetne besedne dele (*pre_*, *o_*), iz tako dobljenih potrjenih glagolov pa po pravilih tvori povezane besede (*pis_at_elj*, *pre_pis_ov_anj_e*, *o_pis_ov_an*). Nekateri tako tvorjeni kandidati so nelegitimni, saj algoritem v trenutni različici pri njih upošteva tudi neustrezne besedne dele (*pis_eti**, *pis_ov_ati**), nekateri pa predstavljajo legitimne enote, ki pa še niso vključene v leksikon (*pre_pis_ov_al_č_ev**,

```
_pis_
     pis_ati
          pis_anj_e  || G.Sos.3.1 || [G]_ati → [G]_anj_e
          pis_at_elj || G.Som.5.2.1 || [G]_ati → [G]_at_elj
               pis_at_elj_ski || Som.P.1.1.1.1 || [S] → [S]_ski
                    pis_at_elj_sk_o || P.R.2.1 || [P]_ski → [P]_sk_o
               pis_at_elj_ev || Som.P.2.2.1 || [S] → [S]_ev
               pis_at_elj_ic_a || Som.Soz.3.1 || [S] → [S]_ic_a
                    pis_at_elj_ič_in || Soz.P.2.1.2 || [S]_ic_a → [S]_ič_in
               [...]
     pis_eti*
     [...]
     pis_ov_ati*
     pre_pis_ov_ati
          pre_pis_ov_anj_e || G.Sos.3.1 || [G]_ati → [G]_anj_e
          pre_pis_ov_al_en || G.P.8.1.1 || [G]_ati → [G]_al_en
          pre_pis_ov_al_ec || G.Som.2.2.1.1 || [G]_ati → [G]_al_ec
               pre_pis_ov_al_č_ev* || Som.P.2.2.2 || [S]_ec → [S]_č_ev
          pre_pis_ov_al_k_a || G.Soz.4.1.1 || [G]_ati → [G]_al_k_a
               pre_pis_ov_al_k_in* || Soz.P.2.1.1 || [S]_a → [S]_in
          [...]
     o_pis_ov_ati
          o_pis_ov_an || G.P.2.1.1 || [G]_ati → [G]_an
          o_pis_ov_al_ec || G.Som.2.2.1.1 || [G]_ati → [G]_al_ec
               o_pis_ov_al_č_ev* || Som.P.2.2.2 || [S]_ec → [S]_č_ev
          o_pis_ov_al_n_ik || G.Som.18.2.1 || [G]_ati → [G]_al_n_ik
          o_pis_ov_al_k_a || G.Soz.4.1.1 || [G]_ati → [G]_al_k_a
               o_pis_ov_al_k_in* || Soz.P.2.1.1 || [S]_a → [S]_in
     [...]
```

Slika 2: Ponazoritev delovanja luščilnega algoritma z glagoli v izhodišču.

*o_pis_ov_al_k_in**). Z dodatnim preverjanjem nenajdenih kandidatov v korpusu (npr. v korpusu pisne standardne slovenščine Gigafida 2.0, s katerim je Sloleks 2.0 povezan) lahko z algoritmom pridobimo tudi nabor potencialnih enot za razširitev leksikona (več o tem v zaključku).

2.3.2 Izhodišče pri samostalnikih, pridevnikih in prislovih

Pri ostalih besednih vrstah, ki so bile vključene v luščenje medbesednih povezav (samostalniki, pridevniki in prislovi), je bil postopek določanja povezav nekoliko manj podroben. Za razliko od glagolov pri ostalih besednih vrstah namreč nismo izhajali iz osrednjih besednih delov, temveč smo kot izhodišče vzeli posamezno iztočnico kot celoto (pri čemer smo preskočili vse iztočnice, ki so bile že obravnavane pri luščenju z glagolskim izhodiščem). Pri vsaki iztočnici smo

preverili, ali se konča na katerega od za njeno besedno vrsto relevantnih končnih besednih delov, jo razcepili (z upoštevanjem najdaljšega možnega končnega dela, npr. *provok_at_or* namesto *provokat_or*, oz. ničtega končnega besednega dela, če ni bilo relevantnega), nato pa na podlagi te delitve na podoben način kot pri luščenju z glagolskim izhodiščem po pravilih rekurzivno tvorili nove kandidate in jih sproti preverjali v leksikonu. Izsek, ki ponazarja delovanje algoritma za luščenje medbesednih povezav z neglagolskim izhodiščem, je prikazan na Sliki 3.

```
faraon
     faraon_ček* || Som.Som.3.1 || [S] → [S]_ček
     faraon_ov  || Som.P.2.1.1 || [S] → [S]_ov
     faraon_ski || Som.P.1.1.1.1 || [S] → [S]_ski
          ne_faraon_ski* || P.P.1 || [P] → ne_[P]
          pre_faraon_ski* || P.P.3 || [P] → pre_[P]
          faraon_sk_o* || P.R.2.1 || [P]_ski → [P]_sk_o
          faraon_s_tvo* || P.Sos.2.1 || [P]_ski → [P]_s_tv_o
     faraon_ov_ec* || Som.Som.1.2.2.1 || [S] → [S]_ov_ec
     [...]
     faraon_k_a || Som.Soz.4.1.1 || [S] → [S]_k_a
          faraon_k_in* || Soz.P.2.1.1 || [S]_a → [S]_in
     faraon_es_a* || Som.Soz.13 || [S] → [S]_es_a
     faraon_j_ad* || Som.Soz.11.1 || [S] → [S]_j_ad
     [...]
```

Slika 3: Ponazoritev delovanja luščilnega algoritma z drugimi besednimi vrstami v izhodišču.

Tudi v tem primeru z algoritmom pridobimo tako kandidate, ki so že vključeni v leksikon (*faraon_ov*, *faraon_ski*, *faraon_k_a*), kot tudi potencialne kandidate za razširitev (*faraon_ček**, *faraon_sk_o**, *faraon_stv_o**, *faraon_k_in**). Za razliko od luščenja z glagoli v izhodišču je treba omeniti, da v tem primeru začetnih besednih delov (oz. njihovih kombinacij) nismo upoštevali, saj so ti pri samostalnikih nekoliko manj predvidljivi kot pri glagolih, osrednjih delov drugih besednih vrst pa nismo ročno pregledali. Tako npr. nismo zajeli medbesednih povezav tipa *soba* → *predsoba*, *škof* → *nadškof*. Izjema sta dve pravili pri povezovanju pridevnikov (npr. *strokoven* → *nestrokoven*, *zadolžen* → *prezadolžen*, glej razdelek 4.3.1), ostala luščenja z začetnimi besednimi deli pa smo pustili za prihodnje delo.

3 Nabor medbesednih povezav

Nabor medbesednih povezav je na voljo na repozitoriju CLARIN.SI (Čibej et al. 2020) v dveh datotekah v formatu TSV: prva vsebuje hierarhijo pravil za vzpostavljanje medbesednih povezav, v drugi pa je navedenih 66.347 edinstvenih izluščenih medbesednih povezav. Datoteka vsebuje izvorno lemo (*abonirati*), povezano lemo (*aboniran*), razcepljeno izvorno lemo (*abon_ir_ati*), razcepljeno povezano lemo (*abon_ir_an*), besedno vrsto izvorne (G) in povezane leme (P), identifikacijski številki obeh lem v Slovenskem oblikoslovnem leksikonu, prekrivni del (*abon*) ter identifikacijsko številko pravila (*G.P.2.1.2*) in povezovalno pravilo (*[G]_ir_ati → [G]_ir_an*). Izsek prikazuje Tabela 5 (zaradi prostorskih omejitev niso prikazani stolpci z nerazcepljenimi lemami in identifikacijskimi številkami iz leksikona).

Tabela 5: Primeri izluščenih medbesednih povezav.

Razcepljena izvorna lema	Razcepljena povezana lema	Besedna vrsta izvorne leme	Besedna vrsta povezane leme	Prekrivni del	ID povezovalnega pravila	Povezovalno pravilo
abon_ir_ati	abon_ir_an	G	P	abon	G.P.2.1.2	[G]_ir_ati → [G]_ir_an
abon_ir_an	abon_ir_an_je	P	Sos	abon	P.Sos.1	[P] → [P]_je
abon_ir_ati	abon_ent	G	Som	abon	G.Som.10	[G]_ir_ati → [G]_ent
abon_ent	abon_ent_ski	Som	P	abon	Som.P.1.1.1.1	[S] → [S]_ski
abon_ent	abon_ent_ov	Som	P	abon	Som.P.2.1.1	[S] → [S]_ov
abon_ent	abon_ent_k_a	Som	Soz	abon	Som.Soz.4.1.1	[S] → [S]_k_a
abon_ir_ati	abon_ma	G	Som	abon	G.Som.19	[G]_ir_ati → [G]_ma
abon_ma	abon_ma_j_ski	Som	P	abon	Som.P.1.1.1.3	[S] → [S]_j_ski
abon_ir_ati	abon_ir_anj_e	G	Sos	abon	G.Sos.3.1	[G]_ati → [G]_anj_e

Kot kaže Tabela 6, je bilo največ povezav izluščenih med pridevniki in občnimi samostalniki ženskega spola (10.101 oz. 15 % vseh povezav), med občnimi samostalniki moškega spola in pridevniki (7.167 oz. slabih 11 %), med glagoli in pridevniki (6.136 oz. dobrih 9 %) ter med pridevniki in prislovi (6.092 oz. dobrih 9 %).

Tabela 6: Število izluščenih povezav v različnih skupinah pravil.

Skupina pravil	Vrsta medbesedne povezave	Število povezav
P.Soz	Povezava med pridevnikom in občnim samostalnikom ženskega spola	10.101
Som.P	Povezava med občnim samostalnikom moškega spola in pridevnikom	7.167
G.P	Povezava med glagolom in pridevnikom	6.136
G.Sos	Povezava med glagolom in občnim samostalnikom srednjega spola	6.092
P.R	Povezava med pridevnikom in prislovom	5.716
Som.Soz	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom ženskega spola	4.755
P.Sos	Povezava med pridevnikom in občnim samostalnikom	4.325
G.Som	Povezava med glagolom in občnim samostalnikom moškega spola	4.116
Soz.P	Povezava med občnim samostalnikom ženskega spola in pridevnikom	4.075
P.Som	Povezava med pridevnikom in občnim samostalnikom moškega spola	2.979
G.Soz	Povezava med glagolom in občnim samostalnikom ženskega spola	2.876
P.P	Povezava med dvema pridevnikoma	2.431
Som.Som	Povezava med dvema občnima samostalnikoma moškega spola	1.087
Soz.Soz	Povezava med dvema občnima samostalnikoma ženskega spola	1.001
G.R	Povezava med glagolom in prislovom	914
Soz.Som	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom moškega spola	826
Sos.P	Povezava med občnim samostalnikom srednjega spola in pridevnikom	816
Som.Sos	Povezava med občnim samostalnikom moškega spola in občnim samostalnikom srednjega spola	586
Soz.Sos	Povezava med občnim samostalnikom ženskega spola in občnim samostalnikom srednjega spola	233
Sos.Sos	Povezava med dvema občnima samostalnikoma srednjega spola	96
Sos.Soz	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom ženskega spola	23
Sos.Som	Povezava med občnim samostalnikom srednjega spola in občnim samostalnikom moškega spola	9

V povprečju so posamezna pravila prispevala približno 122 povezav, polovica več kot 14 povezav. Najmanj produktivna pravila

so doprinesla le po eno povezavo, najproduktivnejše pravilo (*P.R.1* oz. *[P]_en → [P]_n_o*; *hlad_en → hlad_n_o*) pa kar 4.295 povezav. Omeniti je treba, da so bila določena pravila iz hierarhije premalo natančna za luščenje, saj zahtevajo upoštevanje dodatnih pogojev, ki jih v tej različici algoritma še nismo implementirali: to npr. velja za povezovalna pravila iz glagolov, pri katerih je treba za iskanje povezav uporabljati osrednji del sedanjiške oblike namesto nedoločniške (*stre_či – strež_em → strež_aj*). Število pravil, ki so zabeležena v hierarhiji, torej ni nujno enako kot pri luščenju.

4 Evalvacija izluščenih povezav

Da bi preverili, v kolikšni meri so strojno izluščene povezave zanesljive, smo opravili evalvacijo na vzorcu 4.464 povezav, ki so bile vzorčene naključno, a stratificirano po posameznih pravilih (do 10 povezav na pravilo). Povezave smo ročno pregledali in jih označili kot neustrezne, sprejemljive ali ustrezne. Kot ustrezne smo označili povezave, za katere smo presodili (ob upoštevanju Slovenskega etimološkega slovarja in Novega etimološkega slovarja slovenskega jezika, s katerima smo preverili, ali sta besedi morfološko povezani),[5] da bi bile v oblikoslovnem leksikonu glede na besedotvorno povezanost lahko navedene kot povezane iztočnice (npr. *iskati → iskanje*). Kot neustrezne smo označevali povezave, do katerih je prišlo le zaradi naključne površinske podobnosti oblik (npr. *jež → ježa*, *pire → pirejski*). Kot sprejemljive smo označili povezave, ki so sicer do določene mere ustrezne, a pri njih ne gre za neposredno povezavo, temveč za povezavo preko tretje besede (npr. *lasati → lasulja*, obe iztočnici sta v resnici povezani z iztočnico *las*; ustreznost te povezave je sicer odvisna tudi od jezikovnega vira, v katerem se pojavlja, in ali vir od povezav pričakuje samo morfološko ali pa tudi semantično povezanost) oz. za delitev skupnega osrednjega dela, ne pa nujno za neposredno izpeljavo (*sipati → sipina*). Rezultati evalvacije so prikazani v Tabeli 7.

5 Oba slovarja sta dostopna na portalu Fran: https://fran.si/.

Tabela 7: Evalvacija vzorca strojno izluščenih povezav.

Ocena povezave	Število	Delež	Primeri
Ustrezno	3.326	74,51 %	blefirati → blefer topel → toplina datelj → datljev
Sprejemljivo	312	6,99 %	jezikati → jezičen ljubiti → ljubek saditi → sadež
Neustrezno	826	18,50 %	dojeti → doječ pikirati → pikanten plen → plenaren

V nadaljevanju opisujemo podrobnejšo evalvacijo po skupinah pravil glede na besedno vrsto izvorne leme ter izpostavimo najzanesljivejša pravila na eni ter najmanj točna pravila na drugi strani. Omejujemo se le na največ deset najzanesljivejših pravil, v tabelah pa od teh navajamo čimbolj raznovrsten nabor (z različnimi besednimi deli).

4.1 Povezave iz glagolov

Od 4.464 vzorčnih povezav je bila skupno 1.901 povezava (približno 43 % celotnega vzorca) izpeljana neposredno iz glagolskih iztočnic. Evalvacija je predstavljena v Tabeli 8.

Tabela 8: Evalvacija vzorca povezav iz glagolskih iztočnic.

Skupina povezav	Število povezav	Ustrezno		Sprejemljivo		Neustrezno	
G.P	541	428	79 %	48	9 %	65	12 %
G.R	127	114	90 %	0	0 %	13	10 %
G.Som	545	364	67 %	95	17 %	86	16 %
G.Sos	263	227	86 %	4	2 %	32	12 %
G.Soz	425	333	78 %	15	4 %	77	18 %

Glede na analizo vzorca je najvišji delež ustreznih povezav med glagolskimi in prislovnimi iztočnicami (90 %), najnižji pa med glagoli in občnimi samostalniki moškega spola (67 %), pri katerih je v primerjavi z drugimi skupinami tudi nekoliko višji delež sprejemljivih

povezav (17 %). Na nivoju posameznih pravil se pokažejo nekoliko izrazitejše razlike, ki jih predstavljamo v nadaljevanju.

4.1.1 Povezave med glagoli in pridevniki

V vzorcu je pri povezavah med glagoli in pridevniki 26 od 57 pravil doseglo 100-odstotno luščilno točnost (tj. delež povezav, ki smo jih opredelili ko ustrezne; pri tem ne upoštevamo sprejemljivih povezav). Deset od najtočnejših pravil je navedenih v Tabeli 9.

Tabela 9: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in pridevniki.

ID pravila	Pravilo	Primer
G.P.1.1.4	[G]_ati → [G]_aj_oč	naštevati → naštevajoč
G.P.1.1.5	[G]_iti → [G]_uj_oč	gostiti → gostujoč
G.P.1.2.2	[G]_iti → [G]_eč	dušiti → dušeč
G.P.2.1.1	[G]_ati → [G]_an	zvezati → zvezan
G.P.2.3.3	[G]_eti → [G]_et	pregreti → pregret
G.P.3.1	[G]_eti → [G]_el	razvodeneti → razvodenel
G.P.4.7	[G]_ev_ati → [G]_ljiv	obdavčevati → obdavčljiv
G.P.6	[G]_ati → [G]_iv	prebavljati → prebavljiv
G.P.8.1.1	[G]_ati → [G]_al_en	izsiljevati → izsiljevalen
G.P.9	[G]_ir_ati → [G]_abil_en	programirati → programabilen

Od preostalih pravil jih je 16 doseglo vsaj 80-odstotno luščilno točnost, le 7 pravil pa manj kot 50-odstotno točnost (Tabela 10). Nekatera pravila torej niso produktivna oz. dajejo rezultate s precej

Tabela 10: Najmanj točna pravila za povezave med glagoli in pridevniki.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.P.2.2.5	[G]t_iti → [G]č_en	ukrotiti → ukročen	oblatiti → oblačen
G.P.8.2	[G]_eti → [G]_el_en	greti → grelen	streti → strelen
G.P.2.2.7	[G]k_ati → [G]č_en	sekati → sečen	kljukati → ključen
G.P.8.3.2	[G]_ati → [G]_il_en	*barvati → barvilen	razdelati → razdelilen
G.P.5.1	[G]_ati → [G]_ek	*sipati → sipek	šibati → šibek
G.P.5.2	[G]_eti → [G]_ek	*spolzeti → spolzek	trpeti → trpek
G.P.5.3	[G]_iti → [G]_ek	*greniti → grenek	rediti → redek

več šuma kot koristnih povezav: za pravila G.P.5.1, G.P.5.2, G.P.5.3 in G.P.8.3.2 npr. v vzorcu ni bilo niti enega ustreznega primera.

4.1.2 Povezave med glagoli in prislovi

Tudi pri prislovih je večina pravil dosegla 100-odstotno luščilno točnost: od 15 pravil v vzorcu jih je 10 izluščilo samo ustrezne povezave, ostalih 5 pravil pa je doseglo točnost med 50 in 80 %. Najzanesljivejša pravila so prikazana v Tabeli 11.

Tabela 11: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in prislovi.

ID pravila	Pravilo	Primer
G.R.1.1.1	[G]_eti → [G]_eč	šumeti → šumeč
G.R.1.2.1	[G]_ati → [G]_aj_oč	opotekati → opotekajoč
G.R.1.2.2	[G]_ev_ati → [G]_uj_oč	spraševati → sprašujoč
G.R.1.2.3	[G]_ov_ati → [G]_uj_oč	napovedovati → napovedujoč
G.R.1.2.4	[G]_iti → [G]_ij_oč	vpiti → vpijoč
G.R.2.1.1	[G]_ati → [G]_aj_oče	pretakati → pretakajoče
G.R.2.1.4	[G]_iti → [G]_ij_oče	gniti → gnijoče
G.R.2.2.1	[G]_eti → [G]_eče	drveti → drveče
G.R.2.2.4	[G]_iti → [G]_eče	govoriti → govoreče
G.R.2.3	[G]_ati → [G]_aje	vzdihovati → vzdihovaje

Pri povezovanju glagolov in prislovov se je za najmanj zanesljivo izkazalo pravilo G.R.2.4 ([G]_ati → [G]_e), ki je doseglo 50-odstotno točnost: poleg ustreznih kandidatov (*bleščati* → *blešče*, *ležati* → *leže*) je izluščilo tudi precej šumnih povezav, ki so posledica naključne podobnosti oblik (*predati* → *prede*, *divjati* → *divje*). V določenih primerih (npr. *divjati* → *divje*) bi bilo morda smiselno ustrezne povezave od neustreznih ločiti tudi z upoštevanjem naglašenih oblik (*dívje* namesto **divjé*), a Sloleks 2.0 vsebuje le avtomatsko pripisane naglase, ki so manj zanesljivi, poleg tega pa bi bil algoritem, ki se zanaša tudi na naglase, manj primeren za luščenje iz korpusnih oblik, ki so nenaglašene. V prihodnjih različicah leksikona, ki bo vseboval ročno popravljene naglašene oblike, pa bi pri določenih pravilih veljalo upoštevati tudi naglase, vsaj pri postprocesiranju izluščenih povezav.

4.1.3 Povezave med glagoli in občnimi samostalniki

Pri povezavah med glagoli in občnimi samostalniki moškega spola je bilo nekoliko več pravil z nižjo luščilno točnostjo: 23 od 60 pravil je doseglo točnost 60 % ali manj (povprečna točnost je bila 66 %), a je treba upoštevati, da je nekaj od teh pravil večinoma izluščilo sprejemljive povezave – pravilo G.Som.3.2 ([G]_n_iti → [G]; *predahniti → predah*) je npr. izluščilo 90 % sprejemljivih povezav. 10 od 19 najzanesljivejših pravil je prikazanih v Tabeli 12.

Tabela 12: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in občnimi samostalniki moškega spola.

ID pravila	Pravilo	Primer
G.Som.1.1.1	[G]_ati → [G]_aj	cmokljati → cmokljaj
G.Som.1.2.1	[G]_ov_ati → [G]_lj_aj	primanjkovati → primanjkljaj
G.Som.2.2.1.1	[G]_ati → [G]_al_ec	vzdrževati → vzdrževalec
G.Som.2.4.2	[G]_eti → [G]_ev_ec	peti → pevec
G.Som.5.2.2	[G]_iti → [G]_it_elj	voditi → voditelj
G.Som.6.2.3	[G]_iti → [G]_it_ek	dobiti → dobitek
G.Som.7.2.1	[G]_ir_ati → [G]_at_or	likvidirati → likvidator
G.Som.20	[G]_iti → [G]_j_a	voditi → vodja
G.Som.18.2.1	[G]_ati → [G]_al_n_ik	kodrati → kodralnik
G.Som.10	[G]_ir_ati → [G]_ent	abstinirati → abstinent

V Tabeli 13 so prikazana pravila z največjim deležem neustreznih povezav (med 50 in 70 %). Opaziti je mogoče, da do večje količine

Tabela 13: Najmanj točna pravila za povezave med glagoli in občnimi samostalniki moškega spola.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.Som.2.1.3	[G]_iti → [G]_ec	*kriliti → krilec	pobiti → pobec
G.Som.15	[G]_eti → [G]_ez	videti → videz	pogreti → pogrez
G.Som.16.2	[G]d_ir_ati → [G]z_iv	eksplodirati → eksploziv	podirati → poziv
G.Som.5.1.2	[G]_eti → [G]_elj	buhteti → buhtelj	meti → melj
G.Som.13.2	[G]_eti → [G]_uh	smrdeti → smrduh	peti → puh
G.Som.1.1.2	[G]_iti → [G]_aj	enačiti → enačaj	kriti → kraj
G.Som.2.1.1	[G]_ati → [G]_ec	trgovati → trgovec	zmajevati → zmajevec

neustreznih povezav pride pri nekoliko bolj specifičnih pravilih, pri katerih je poleg končnega besednega dela upoštevan tudi del osrednjega besednega dela (npr. *eksplodirati → eksploziv*), in pri pravilih, ki vključujejo manj produktivne končne besedne dele (npr. *vid_ez*), zaradi česar pravilo pogosteje zajame oblike, ki se po naključju končajo na enako zaporedje črk (*pogreti → pogrez*).

Pri povezavah med glagoli in občnimi samostalniki ženskega spola je bila povprečna luščilna točnost višja (79 %) kot pri samostalnikih moškega spola. 22 od 52 pravil je izluščilo samo ustrezne povezave (10 jih je naštetih v Tabeli 14), še 16 pravil pa je doseglo nadpovprečno točnost (med 80 in 94 %).

Tabela 14: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in občnimi samostalniki ženskega spola.

ID pravila	Pravilo	Primer
G.Soz.1.1.1	[G]_ir_ati → [G]_ac_ij_a	migrirati → migracija
G.Soz.1.2.2	[G]n_ir_ati → [G]z_ic_ij_a	komponirati → kompozicija
G.Soz.1.4.1	[G]h_ir_ati → [G]kc_ij_a	abstrahirati → abstrakcija
G.Soz.11.3	[G]_lj_ati → [G]_a	zlorabljati → zloraba
G.Soz.12.1.1	[G]_ati → [G]_il_j_a	šivati → šivilja
G.Soz.14.2	[G]z_iti → [G]ž_nj_a	groziti → grožnja
G.Soz.2	[G]_ati → [G]_ar_ij_a	pisati → pisarija
G.Soz.3.1	[G]_ati → [G]_at_ev	dajati → dajatev
G.Soz.4.1.1	[G]_ati → [G]_al_k_a	izpraševati → izpraševalka
G.Soz.9.1	[G]_iti → [G]_b_a	obeležiti → obeležba

Osem pravil je doseglo luščilno točnost pod 50 % (Tabela 15). Tudi v teh primerih je vzrok za neustrezne povezave največkrat naključna podobnost oblik (*pomirati → pomada*, *stepsti → stepa*), v primeru pravila G.Soz.15 pa gre za zelo redek končni besedni del (*_uša*).

Za povezave med glagoli in občnimi samostalniki srednjega spola je bilo v vzorcu manj pravil, skupno 29 s povprečno luščilno točnostjo približno 80 %. Pri 14 pravilih so bile vse evalvirane povezave ustrezne, pri še petih pa je bila točnost nadpovprečna (med 84 in 92 %). Deset od najzanesljivejših pravil v tej skupini je navedenih v Tabeli 16.

Tabela 15: Najmanj točna pravila za povezave med glagoli in občnimi samostalniki ženskega spola.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.Soz.8	[G]_ir_ati → [G]_ad_a	blokirati → blokada	pomirati → pomada
G.Soz.11.6	[G]_ov_ati → [G]_a	prevladovati → prevlada	kupovati → kupa
G.Soz.11.5	[G]_sti → [G]_a	pozebsti → pozeba	stepsti → stepa
G.Soz.15	[G]_eti → [G]_uš_a	poleteti → poletuša	deti → duša
G.Soz.6.3	[G]_iti → [G]_j_av_a	*širiti → širjava	tuliti → tuljava
G.Soz.6.2	[G]_iti → [G]_av_a	težiti → težava	ustiti → ustava
G.Soz.11.4	[G]_eti → [G]_a	oskrbeti → oskrba	pričeti → priča
G.Soz.6.4	[G]_iti → [G]_nj_av_a	*bloditi → blodnjava	motiti → motnjava

Tabela 16: Deset pravil s 100-odstotno točnostjo za povezave med glagoli in občnimi samostalniki srednjega spola.

ID pravila	Pravilo	Primer
G.Sos.1.3	[G]_iti → [G]_išč_e	gnezditi → gnezdišče
G.Sos.1.4	[G]_ati → [G]_al_išč_e	pristajati → pristajališče
G.Sos.2.1	[G]_ati → [G]_al_o	gobezdati → gobezdalo
G.Sos.2.5	[G]_iti → [G]_il_o	beliti → belilo
G.Sos.3.1	[G]_ati → [G]_anj_e	razdirati → razdiranje
G.Sos.3.2.3	[G]_iti → [G]_j_enj_e	žepariti → žeparjenje
G.Sos.3.3.1	[G]_iti → [G]_en_je	obeležiti → obeleženje
G.Sos.3.4.1	[G]_eti → [G]_et_je	najeti → najetje
G.Sos.3.4.2	[G]_iti → [G]_it_je	izliti → izlitje
G.Sos.4.1	[G]_iti → [G]_iv_o	razstreliti → razstrelivo

V Tabeli 17 so navedena pravila z najvišjim deležem neustreznih povezav. Zanimivo je, da gre pri vseh manj zanesljivih pravilih za podskupine oz. podpravila najbolj zanesljivih pravil: luščenje povezav s samostalniki s končnim besednim delom *_iv_o* je npr. zelo zanesljivo pri glagolih s končnim besednim delom *_iti* (G.Sos.4.1 iz Tabele 17), a precej manj zanesljivo pri glagolih na *_ati* (G.Sos.4.3) in *_eti* (G.Sos.4.2).

Tabela 17: Najmanj točna pravila za povezave med glagoli in občnimi samostalniki srednjega spola.

ID pravila	Pravilo	Ustrezen (ali *sprejemljiv) primer	Neustrezen primer
G.Sos.2.3	[G]e_sti → [G]_el_o	omesti → omelo	sesti → selo
G.Sos.3.4.5	[G]_iti → [G]_ot_je	ganiti → ganotje	priti → protje
G.Sos.4.3	[G]_ati → [G]_iv_o	mazati → mazivo	predati → predivo
G.Sos.1.5	[G]_eti → [G]_el_išč_e	vreti → vrelišče	streti → strelišče
G.Sos.1.7	[G]_n_iti → [G]_l_išč_e	zmrzniti → zmrzlišče	meniti → melišče
G.Sos.3.4.4	[G]_iti → [G]_ut_je	preminiti → preminutje	počiti → počutje
G.Sos.4.2	[G]_eti → [G]_iv_o	goreti → gorivo	peti → pivo

4.2 Povezave iz občnih samostalnikov

Povezave iz občnih samostalnikov v vzorcu zajemajo približno 49 % (2.191 povezav). Evalvacija luščilne točnosti je predstavljena v Tabeli 18. V povprečju so pravila dosegla 77 % točnost. Najvišjo točnost lahko opazimo pri povezavah med občnimi samostalniki srednjega spola in drugimi občnimi samostalniki (med 91 in 100 %). Nekoliko manj zanesljive so povezave med občnimi samostalniki ženskega spola in ostalimi občnimi samostalniki (do 33 % neustreznih povezav).

Tabela 18: Evalvacija vzorca povezav iz samostalniških iztočnic.

Skupina povezav	Število povezav	Ustrezno		Sprejemljivo		Neustrezno	
Som.P	549	401	73 %	10	2 %	138	25 %
Som.Som	232	167	72 %	9	3 %	56	25 %
Som.Sos	68	52	76 %	1	1 %	15	23 %
Som.Soz	191	133	70 %	13	7 %	45	23 %
Sos.P	87	57	66 %	24	27 %	6	7 %
Sos.Som	9	9	100 %	0	0 %	0	0 %
Sos.Sos	23	21	91 %	0	0 %	2	9 %
Sos.Soz	23	23	100 %	0	0 %	0	0 %
Soz.P	464	347	75 %	20	4 %	97	21 %
Soz.Som	320	193	60 %	22	7 %	105	33 %
Soz.Sos	86	57	66 %	11	13 %	18	21 %
Soz.Soz	139	98	71 %	12	8 %	29	21 %

4.2.1 Povezave med občnimi samostalniki in pridevniki

Pri občnih samostalnikih moškega spola je bilo v vzorcu kar 72 pravil za povezave s pridevniki, v povprečju pa je bila njihova luščilna točnost 73-odstotna. 31 pravil je izluščilo samo ustrezne povezave, še 13 pa jih je bilo nadpovprečno točnih. Deset od najzanesljivejših pravil je prikazanih v Tabeli 19.

Tabela 19: Deset pravil s 100-odstotno točnostjo za povezave med občnimi samostalniki moškega spola in pridevniki.

ID pravila	Pravilo	Primer
Som.P.1.1.1.1	[S] → [S]_ski	čolnar → čolnarski
Som.P.1.1.2.4	[S]er → [S]r_ov_ski	kader → kadrovski
Som.P.1.1.3.2	[S]_ec → [S]_č_ev_ski	borec → borčevski
Som.P.1.2.3	[S]š → [S]_ški	bogataš → bogataški
Som.P.2.1.1	[S] → [S]_ov	tat → tatov
Som.P.3.1.5	[S]er → [S]r_n	alabaster → alabastrn
Som.P.3.7.1.4	[S]_ek → [S]_k_ov_en	podatek → podatkoven
Som.P.4.6	[S]_ec → [S]_č_ast	apnenec → apnenčast
Som.P.5.7	[S]_ec → [S]_č_ji	zajec → zajčji
Som.P.6.2.2	[S]_ek → [S]_k_ov_it	učinek → učinkovit

Med najbolj problematičnimi pravili so Som.P.1.1.3.3 ([S] → [S]_j_ev_ski), Som.P.1.1.5 ([S] → [S]_j_an_ski), Som.P.5.5 ([S]k → [S] č_ji), Som.P.6.1.1 ([S] → [S]_it) in Som.P.9.2.2 ([S]g → [S]ž_n_at), ki v vzorcu niso imeli niti ene ustrezne povezave (npr. *bar* → *barjanski*, *pob* → *pobit*, *rak* → *račji*, *rog* → *rožnat*). Pri teh je treba preveriti vzrok za slabe rezultate (npr. napaka v luščilnem algoritmu ali pravilu) in pravila po potrebi prilagoditi ali odstraniti iz hierarhije oz. iz luščilnega postopka.

Pravil za povezovanje občnih samostalnikov srednjega spola in pridevnikov je bilo v vzorcu 14, od tega jih je 9 izluščilo samo ustrezne povezave (Tabela 20).

Pri ostalih pravilih je točnost nekoliko nižja, a je neustreznih povezav kljub temu malo (do 15 %). Preostale povezave so sprejemljive, npr. pri pravilu Sos.P.1 ([S]_o → [S]_ski, *vino* → *vinski*), kjer zaradi podobnosti končnih besednih delov prihaja do prekrivnosti z

drugimi pravili (npr. *kadilo* → *kadilski*, kjer bi bila ustreznejša povezava *kadilec* → *kadilski*).

Tabela 20: Pravila s 100-odstotno točnostjo za povezave med občnimi samostalniki srednjega spola in pridevniki.

ID pravila	Pravilo	Primer
Sos.P.2	[S]c_e → [S]č_ev	sonce → sončev
Sos.P.3.1.2	[S]_e → [S]_en	razstavišče → razstaviščen
Sos.P.3.1.5	[S]k_o → [S]č_en	jabolko → jabolčen
Sos.P.3.2.1	[S]r_o → [S]r_n	jedro → jedrn
Sos.P.3.2.2	[S]l_o → [S]el_n	sedlo → sedeln
Sos.P.3.3	[S]_o → [S]_ov_en	delo → deloven
Sos.P.9.1	[S]_o → [S]_n_at	meso → mesnat
Sos.P.9.2	[S]k_o → [S]č_n_at	mleko → mlečnat
Sos.P.9.3	[S]_e → [S]_n_at	olje → oljnat

Od 52 pravil za povezovanje občnih samostalnikov ženskega spola s pridevniki je bilo 23 100-odstotno točnih (povprečna luščilna točnost je bila 78 %), 10 od teh jih je prikazanih v Tabeli 21.

Tabela 21: Deset pravil s 100-odstotno točnostjo za povezave med občnimi samostalniki srednjega spola in pridevniki.

ID pravila	Pravilo	Primer
Soz.P.1.1	[S]_a → [S]_ski	lokacija → lokacijski
Soz.P.2.1.1	[S]_a → [S]_in	oškodovanka → oškodovankin
Soz.P.3.1.2	[S] → [S]_en	težnost → težnosten
Soz.P.3.1.4	[S]c_a → [S]č_en	lestvica → lestvičen
Soz.P.3.2	[S]_ev → [S]_v_en	meritev → meritven
Soz.P.3.6.1.2	[S]_ij_a → [S]_iv_en	korozija → koroziven
Soz.P.3.6.2.1	[S]_ac_ij_a → [S]_at_iv_en	provokacija → provokativen
Soz.P.4.1.1	[S]_a → [S]_ast	krogla → kroglast
Soz.P.5.1.4	[S]c_a → [S]č_ji	veverica → veverički
Soz.P.9.2.2	[S]k_a → [S]č_n_at	opeka → opečnat

Med najbolj problematičnimi pravili (z več kot 50 % neustreznimi povezavami) so Soz.P.1.2 ([S]_a → [S]_ov_ski, *peka* → *pekovski*), Soz.P.3.1.6 ([S]_ij_a → [S]_en, *alotropija* → *alotropen*), Soz.P.3.7.1 ([S]_a → [S]_ov_en, *cena* → *cenoven*), in Soz.P.3.3.1 ([S]_a → [S]_ič_en,

metafora → metaforičen). Verjetno je, da so pravila, ki izpeljujejo povezave iz besed z zelo splošnimi in pogostimi končnimi besednimi deli (npr. *_a*), nekoliko bolj podvržena naključnemu šumu.

4.2.2 Povezave med občnimi samostalniki

V vzorcu je vseh pravil za povezave med različnimi kombinacijami občnih samostalnikov ženskega, srednjega in moškega spola skupno 142. V tem razdelku se zaradi prostorskih omejitev osredotočamo le na nekatere od tistih, ki so izluščili največ povezav znotraj svoje kategorije (Tabela 22).

Tabela 22: Najproduktivnejša pravila za povezave med občnimi samostalniki ženskega, srednjega in moškega spola.

ID pravila	Pravilo	Luščilna točnost	Ustrezen primer	Neustrezen primer
Som.Som.1.1	[S] → [S]_ec	90 %	duh → duhec	bor → borec
Som.Som.3.1	[S] → [S]_ček	90 %	kurir → kurirček	kov → kovček
Som.Som.3.2	[S]_ec → [S]_ček	100 %	vesoljec → vesoljček	/
Som.Som.15	[S]_izem → [S]_ist	100 %	absolutizem → absolutist	/
Som.Sos.2.1	[S] → [S]_stv_o	90 %	vohun → vohunstvo	roj → rojstvo
Som.Sos.3.1	[S] → [S]_išč_e	100 %	prizor → prizorišče	/
Som.Soz.1.1	[S] → [S]_a	60 %	soprog → soproga	por → pora
Som.Soz.3.1	[S] → [S]_ic_a	90 %	ravnatelj → ravnateljica	krst → krstica
Som.Soz.4.1.1	[S] → [S]_k_a	90 %	recenzent → recenzentka	govor → govorka
Som.Soz.4.1.2	[S]_ec → [S]_k_a	100 %	tvorec → tvorka	/
Sos.Sos.1	[S]_o → [S]_ce	100 %	besedilo → besedilce	/
Soz.Som.1.1	[S]_a → [S]_ec	60 %	kmetija → kmetijec	soda → sodec
Soz.Som.16.2	[S]c_ij_a → [S]t_or	100 %	ilustracija → ilustrator	/
Soz.Sos.1.1	[S]_a → [S]_je	60 %	beseda → besedje	peta → petje
Soz.Soz.1.1	[S]_a → [S]_ic_a	80 %	naprava → napravica	lisa → lisica
Soz.Soz.1.4	[S]_a → [S]_n_ic_a	90 %	zaščita → zaščitnica	nakaza → nakaznica

Večina najproduktivnejših pravil za povezave med občnimi samostalniki je pri evalvaciji dosegla visoko točnost (90 oz. 100 %), največji delež neustreznih povezav pa so imela pravila Som.Soz.1.1 (*soprog → soproga*), Soz.Som.1.1 (*kmetija → kmetijec*) in Soz.Sos.1.1 (*beseda → besedje*) – tudi pri teh se kaže, da je problematičen končni besedni del *_a*, ki privede do precejšnje mere šuma zaradi površinske podobnosti oblik (*por → pora*, *jež → ježa*). Na drugi strani so zelo regularna in zanesljiva nekatera pravila s končnimi besednimi deli latinskega izvora (*ilustracija → ilustrator*, *absolutizem → absolutist*) ter s pari končnih besednih delov, ki nekoliko bolj nedvoumno povezujejo relevantne iztočnice (*vesoljec → vesoljček*, *besedilo → besedilce*, *tvorec → tvorka*). Pri pravilu Soz.Soz.1.4 (*zaščita → zaščitnica*) se pojavi vprašanje, kako obravnavati povezave, ki jih lahko s pravili vzpostavimo na več načinov (npr. *zaščita → zaščitnica*, *zaščita → zaščiten → zaščitnica*). To z vidika samih povezav med iztočnicami, kot so podane v leksikonu, ni tako problematično, terja pa dodaten premislek za morebitno gradnjo morfoloških derivacijskih dreves, pri katerih so besede razporejene v hierarhijo.

4.3 Povezave iz pridevnikov

V evalviranem vzorcu predstavljajo povezave iz pridevnikov le približno 8 % (skupno 372 povezav), največ povezav pa je z občnimi samostalniki ženskega spola. Evalvacijo povezav po skupinah prikazuje Tabela 23. V treh skupinah pravil v vzorcu ni bilo neustreznih povezav, le pri povezavah z občnimi samostalniki ženskega in moškega spola jih je bil manjši delež (14 in 17 %).

Tabela 23: Evalvacija vzorca povezav iz pridevniških iztočnic.

Skupina povezav	Število povezav	Ustrezno		Sprejemljivo		Neustrezno	
P.P	35	31	89 %	4	11 %	0	0 %
P.R	30	30	100 %	0	0 %	0	0 %
P.Som	89	64	72 %	10	11 %	15	17 %
P.Sos	30	20	67 %	10	33 %	0	0 %
P.Soz	188	157	84 %	4	2 %	27	14 %

4.3.1 Povezave med dvema pridevnikoma ter pridevniki in prislovi

Pravil za povezave med dvema pridevnikoma ter pridevniki in prislovi je v vzorcu zgolj 8 (5 za P.P in 3 za P.R), zato skupini obravnavamo skupaj in vsa pravila naštevamo v Tabeli 24. Rezultati potrjujejo, da so povezave med pridevniki in prislovi zelo regularne. Edino pravilo, ki ni doseglo 100-odstotne luščilne točnosti, je P.P.3, pri katerem je večina povezav z elativom (*lep* → *prelep*), nekatere povezave pa so zgolj sprejemljive (npr. *vozniški* → *prevozniški*).

Tabela 24: Pravila za povezave med dvema pridevnikoma oz. med pridevnikom in prislovom.

ID pravila	Pravilo	Primer
P.P.1	[P] → ne_[P]	strokoven → nestrokoven
P.P.3	[P] → pre_[P]	zadolžen → prezadolžen
P.P.4.1	[P] → [P]_ik_ast	črn → črnikast
P.P.4.2	[P] → [P]_k_ast	slan → slankast
P.P.5	[P] → [P]_lj_at	gost → gostljat
P.R.1	[P]_en → [P]_n_o	kritičen → kritično
P.R.2.1	[P]_ski → [P]_sk_o	vrhunski → vrhunsko
P.R.2.2	[P]_ški → [P]_šk_o	geološki → geološko

4.3.2 Povezave med pridevniki in občnimi samostalniki

V vzorcu je povezovalnih pravil med pridevniki in občnimi samostalniki moškega spola 14, od teh jih je 9 izluščilo samo ustrezne povezave (Tabela 25). Omeniti je treba, da so nekatera pravila – npr. P.Som.11, P.Som.4.2, P.Som.9 in P.Som.8.2 – vezana na precej majhen nabor iztočnic (*beluš*, *modrijan*, *lenuh*/*debeluh*/*skopuh*, *mrtvak*) in so za nadaljnje luščenje povezav manj primerna, druga pa so mnogo bolj produktivna (npr. P.Som.1.1 in P.Som.12).

Najmanj točni sta bili sicer nizkoproduktivni pravili P.Som.8.3 ([P] → [P]_ak, *prost* → *prostak*, 50 % neustreznih povezav, npr. *kul* → *kulak*) in P.Som.7 ([P] → [P]_k_ar, *rdeč* → *rdečkar*, 80 % neustreznih povezav, npr. *križan* → *križankar*).

Tabela 25: Pravila s 100-odstotno točnostjo za povezave med pridevniki in občnimi samostalniki moškega spola.

ID pravila	Pravilo	Primer
P.Som.1.1	[P] → [P]_ec	razseljen → razseljenec
P.Som.10	[P] → [P]_un	čist → čistun
P.Som.11	[P] → [P]_uš	bel → beluš
P.Som.12	[P]_en → [P]_n_ik	dvomesečen → dvomesečnik
P.Som.2.1	[P] → [P]_ež	ognjevit → ognjevitež
P.Som.2.2	[P]_en → [P]_n_ež	izviren → izvirnež
P.Som.4.2	[P]er → [P]r_ij_an	moder → modrijan
P.Som.8.2	[P]ev → [P]v_ak	mrtev → mrtvak
P.Som.9	[P] → [P]_uh	len → lenuh

Za povezave med pridevniki in občnimi samostalniki srednjega spola so v vzorcu le tri pravila: P.Sos.2.1 ([P]_ski → [P]_s_tvo, *bibliotekarski → bibliotekarstvo*) in P.Sos.2.2 ([P]_ški → [P]_š_tvo, *zarotniški → zarotništvo*) sta izluščila le ustrezne povezave. Povezave, izluščene s pravilom P.Sos.1 ([P] → [P]_je, *ocvetličen → ocvetličenje*), smo pri evalvaciji označili za sprejemljive – pravilo je namreč deloma prekrivno z določenimi pravili za povezave med glagoli in občnimi samostalniki (*ocvetličiti → ocvetličenje*), zato je potreben dodaten premislek, ali eno od pravil iz hierarhije odstranimo.

Tabela 26: Deset pravil s 100-odstotno točnostjo za povezave med pridevniki in občnimi samostalniki ženskega spola.

ID pravila	Pravilo	Primer
P.Soz.1.1	[P] → [P]_k_a	domišljav → domišljavka
P.Soz.10	[P] → [P]_ul_j_a	kosmat → kosmatulja
P.Soz.11	[P]_iv_en → [P]_iv_a	perspektiven → perspektiva
P.Soz.12	[P] → [P]_oč_a	nečist → nečistoča
P.Soz.3.1.2	[P]_en → [P]_n_in_a	donosen → donosnina
P.Soz.3.2.1	[P]_er → [P]_r_ič_in_a	dober → dobričina
P.Soz.3.3.1	[P]_ski → [P]_šč_in_a	portugalski → portugalščina
P.Soz.4	[P]_en → [P]_n_j_av_a	bloden → blodnjava
P.Soz.5	[P]_en → [P]_n_ic_a	dvozložen → dvozložnica
P.Soz.7.1.1	[P] → [P]_ost	razčlenjen → razčlenjenost

Kar 19 od 27 povezovalnih pravil med pridevniki in občnimi samostalniki ženskega spola je bilo 100-odstotno točnih (deset jih je prikazanih v Tabeli 26). Najbolj produktivna pravila so P.Soz.7.1.1, P.Soz.5, P.Soz.3.3.1 in P.Soz.1.1, po obsegu zelo omejeni pravili pa sta npr. P.Soz.12 in P.Soz.3.2.1.

Problematična so le tri pravila, ki so izluščila med 45 in 60 % neustreznih primerov: P.Soz.6.1.4 ([P]st_en → [P]šč_ob_a), ki je izluščil le dve povezavi: *masten* → *maščoba* in neustrezno povezavo *pusten* → *puščoba*; P.Soz.6.1.2 ([P]_en → [P]_ob_a; ustrezen primer je *gnusen* → *gnusoba*, med neustreznimi pa sta npr. *poden* → *podoba* in *milen* → *miloba*) in P.Soz.8.1.2 ([P]_en → [P]_ot_a, *grozen* → *grozota*, a neustrezno *siren* → *sirota*).

4 Sklep

V prispevku smo predstavili prvi korak k strojnemu luščenju medbesednih povezav v oblikoslovnem leksikonu Sloleks. V primerjavi z različico 2.0, ki vsebuje 30.502 edinstveni medbesedni povezavi (brez upoštevanja lastnoimenskih samostalnikov), smo z robustno metodo na podlagi povezovalnih pravil izluščili 66.347 edinstvenih medbesednih povezav. Preliminarna evalvacija kaže, da je metoda uspešna, saj so tako pridobljene povezave v povprečju zanesljive v približno 75–80 % primerov (odvisno od pravila). Poleg medbesednih povezav, s katerimi bo mogoče dopolniti leksikon, je rezultat raziskave tudi prva različica odprto dostopne baze s strojno berljivimi podatki o slovenskem besedotvorju, ki vsebuje formalizirana in robustna povezovalna besedotvorna pravila, prilagojena avtomatski obdelavi naravnega jezika.

V prihodnje bi bilo smiselno hierarhijo povezovalnih pravil dopolniti z dodatnimi pravili z upoštevanjem lastnih imen (*Novak_ov*, *godovi_ški*) in delov zloženk kot besednih delov (*hidro_elektr_arn_a*), kar smo v trenutnem luščilnem postopku preskočili. Izvesti bi bilo treba tudi natančnejšo in obsežnejšo evalvacijo povezovalnih pravil, saj je trenutna evalvacija temeljila na relativno majhnem vzorcu (do 10 povezav na pravilo). Obsežnejša evalvacija bi pomagala odstraniti

šum, pridobljen s strojnim luščenjem, omogočila pa bi tudi jasnejšo kvantifikacijo produktivnosti in zanesljivosti posameznih pravil.

Potrebne so tudi določene izboljšave znotraj obstoječih pravil in povezav: povezati je npr. treba dovršne in nedovršne glagole (npr. *ugotoviti – ugotavljati*), ki so trenutno v primerih, ko se osrednji del razlikuje med različnimi oblikami, obravnavani ločeno, zaradi česar ne dobimo povezave *ugotoviti → ugotavljanje*. Podobno je treba izboljšati tudi luščenje npr. iz glagolov na *_sti – jesti → jedec*, *pregristi → pregriznjen*. Obenem je treba dodati tudi pravila, s katerimi druge besedne vrste povezujemo z glagoli – v trenutni različici smo zaradi načina luščilnega algoritma glagole vedno obravnavali kot izhodiščne, četudi nekateri izhajajo iz drugih besednih vrst, npr. *urad → uradovati*, *predsednik → predsednikovati*, *rumen → rumeneti*).

V okviru oblikoslovnega leksikona je treba določiti kriterije, po katerih so navedene povezane iztočnice, in razdvoumiti razlike med (zgolj) morfološko sorodnimi (*plamen → plamenec*) in (tudi) semantično sorodnimi pari (*tekmovati → tekmovalec*). To zadeva pomembno in splošnejše vprašanje, kako je v Sloleksu obravnavan pomen, v okviru tega pa je treba razrešiti še nekatere druge dileme – v različici 2.0 npr. niso ločene iztočnice po naglasih, ki razlikujejo pomen (npr. *drèn – drén*).

Ker algoritem kot rezultat pravil ponudi tudi kandidate, ki še niso vključeni v leksikon (razdelek 2.3), bi bilo smiselno metodo preizkusiti tudi za iskanje kandidatov za razširjanje leksikona v korpusih, kot je korpus pisne standardne slovenščine Gigafida. Postopek evalvacije izluščenih povezav bi se potencialno lahko uporabil tudi v slovaropisnem postopku, saj leksikograf_inja lahko dobi seznam kandidatov za povezana gesla, ki jih izbere, s tem pa hkrati opremlja tudi oblikoslovni leksikon.

Določiti bi bilo treba tudi strojno berljive besednodelitvene vzorce glede na besednodelno strukturo (npr. *na_pis_ati* → [začetni]-[osrednji]-[končni], *o_pis_ov_ati* → [začetni]-[osrednji]-[končni]-[končni]) ter generirati derivacijsko morfološko mrežo za slovenščino, kar bi še dodatno dopolnilo jezikovno opremljenost slovenščine v digitalni dobi.

Zahvala

Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (šifra ARRS: J6-8256) in raziskovalni program št. P6-0411 – Jezikovni viri in tehnologije za slovenščino je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Avtor se zahvaljuje Evi Pori za pomoč pri pripravi nabora besednih delov na podlagi Slovenske slovnice, Miji Bon za pomoč pri pregledu luščenja iz Sloleksa na podlagi predpon in ekipi projekta NSSSS za posvetovanje pri pisanju povezovalnih pravil.

Reference

Čibej, J., Arhar Holdt, Š. in Krek, S. (2020). List of word relations from the Sloleks 2.0 lexicon 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1386.

Dobrovoljc, K., Krek, S. in Erjavec, T. (2015). Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, Gantar, P., Kosem, I. in Krek, S. (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 80–105). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/489-1.

Kern, B. (2010). Stopenjsko besedotvorje. *Slavistična revija*, 58 (3), 335–348. Dostopno prek: https://srl.si/ojs/srl/article/view/COBISS_ID-31807533.

Kern, B. (2011). Analiza besedotvornih sklopov glagola stopiti. *Jezikoslovni zapiski,* 17, 127–141. Dostopno prek: https://ojs.zrc-sazu.si/jz/issue/view/206.

Kern, B. (2017). *Stopenjsko besedotvorje. Na primeru glagolov čutnega zaznavanja*. Ljubljana: Založba ZRC. https://doi.org/10.3986/9789610504191.

Kern, B. (2020). Kombinatorika priponskih obrazil v besedotvornih sestavih glagolov čutnega zaznavanja. V M. Kranjc Ivič in A. Žele (ur.), *Pogled v jezik in iz jezika: Adi Vidovič Muha ob jubileju* (str. 67–79). Maribor: Univerzitetna založba. https://doi.org/10.18690/978-961-286-334-0.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020:*

Twelfth International Conference on Language Resources and Evaluation: conference proceedings (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Lango, M., Ševčíková, M. in Žabokrtský, Z. (2018). Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). V N. Calzolari et al. (ur.), *LREC 2018: Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (str. 1853–1860). Pariz: European Language Resources Association. Dostopno prek: https://aclanthology.org/volumes/L18-1/.

Lignos, C., Chan E., Marcus, M. P. in Yang, C. (2009). A rule-based unsupervised morphology learning framework. *Working Notes for the CLEF 2009 Workshop*. Dostopno prek: http://ceur-ws.org/Vol-1175/CLEF2009wn-MorphoChallenge-LignosEt2009.pdf.

Jakopin, F. (1971). Glagoli premikanja v slovenščini in ruščini. V J. Toporišič (s sodelovanjem Alenke Logar Pleško) (ur.), *VII. seminar slovenskega jezika, literature in kulture, 5.–17. julij 1971* (str. 12). Ljubljana: Filozofska fakulteta, Oddelek za slovanske jezike in književnosti.

Jakopin, P., Michelizza, M. in Žele, A. (2009). Besedotvorne smernice v slovenščini v okviru predponskoobrazilnih tvorjenk in zloženk. V. A. Gložančev et al. (ur.), *Novejša slovenska leksika: v povezavi s spletnimi jezikovnimi viri* (str. 203–409). Ljubljana: Založba ZRC. https://doi.org/10.3986/9789610503927.

Skarżyński, M. (2000). *Liczebniki w słowotwórstwie współczesnej polszczyzny (Studium gniazd słowotwórczych)*. Krakov: Towarzystwo Wydawnicze »Historia Iagellonica«.

Stramljič Breznik, I. (2020). *Besedotvorje: teoretično, praktično in didaktično.* Maribor: Univerzitetna založba Univerze v Mariboru. https://doi.org/10.18690/978-961-286-380-7.

Ševčíková, M. (2018). Modelling Morphographemic Alternations in Derivation of Czech. *The Prague Bulletin of Mathematical Linguistics,* 110*,* 7–42. Dostopno prek: https://ufal.mff.cuni.cz/pbml/110/art-sevcikova.pdf.

Vidovič Muha, A. (1988). *Slovensko skladenjsko besedotvorje ob primerih zloženk*. Ljubljana: Znanstvena založba Filozofske fakultete.

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

Zeller, B., Šnajder, J. in Padó, S. (2013). DERIVBASE: Inducing and Evaluating a Derivational Morphology Resource for German. V H. Schuetze et al. (ur.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (str. 1201–1211). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/P13-1118.pdf.