

Jaka Čibejⁱ

Empirična analiza naglašanih enot v slovenščini na podlagi *Slovenskega oblikoslovnega leksikona Sloleks*

V prispevku predstavljamo postopek gradnje empirično podprtega izčrpnega seznama naglašanih enot v slovenščini na podlagi *Slovenskega oblikoslovnega leksikona Sloleks*, in sicer s poudarkom na stalno naglašanih enotah. Obstoječi jezikovni priročniki (npr. *Slovenska slovnica*, *Slovenski pravopis* 2001) namreč navajajo le delne podatke brez navedenih pogostosti, strojno berljiv seznam stalno naglašanih enot pa bi zaradi le deloma predvidljivega naglasnega mesta v slovenščini koristil ne le jezikovnim govorcem, temveč tudi strojnemu naglaševalniku. Iz leksikona najprej izvozimo vse naglašene enote in za vsakega od izvoženih naglašanih besednih delov preštejemo, v koliko leksemih se pojavlja, ter izračunamo deleže, v katerih je del naglašen. Na tako pridobljenih podatkih opravimo analizo po besednih vrstah in primerjamo razlike med rezultati, pridobljenimi po našem postopku, in obstoječimi jezikovnimi priročniki. Opišemo sestavo tako nastalega seznama, ki je odprto dostopen na repozitoriju CLARIN.SI, in dokumentiramo postopek njegove gradnje. Novo različico seznama je na enak način mogoče ob posodobitvah leksikona izvoziti z upoštevanjem novih podatkov, ki natančneje odsevajo podobo naglaševanja v slovenščini. V zaključku navedemo še nekaj možnosti za izboljšave in potencialne korake za nadaljnje raziskave.

Ključne besede: stalno naglašene enote, naglaševanje, slovenščina, oblikoslovni leksikon, Sloleks

An Empirical Analysis of Accentuated Units in Slovene Based on the *Sloleks Morphological Lexicon of Slovene*

In this paper, we present the process of compiling an empirically based comprehensive list of accentuated units in Slovene based on the *Sloleks Morphological Lexicon of Slovene*, with an emphasis on constantly

ⁱ Univerza v Ljubljani, Filozofska fakulteta in Center za jezikovne vire in tehnologije; jaka.cibej@ff.uni-lj.si; <https://orcid.org/0000-0002-3037-6848>

accentuated units. Existing language manuals (e.g. the *Slovene Grammar* and the *Slovenian Normative Guide* 2001) provide only partial data without specified frequencies. A machine-readable list of constantly accentuated units would be useful not only to language speakers, but also for automatic accentuators due to the only partly predictable place of accentuation in Slovene. We first export all accentuated units from the lexicon, then for each of the exported accentuated word parts we count how many lexemes it occurs in. We calculate the proportions in which the relevant word part is accentuated. We analyse the data obtained in this way by word types and compare the differences between our results and existing Slovene language manuals. We describe the structure of the resulting list, which is openly accessible in the CLARIN.SI repository, and document the process of its compilation. Using the same process, a new version of the list can always be exported when the lexicon is updated, taking into account new data in order to accurately reflect accentuation in Slovene. We finish the paper by listing several possibilities for improvement and potential steps for further research.

Keywords: constantly accentuated units, accentuation, Slovene, morphological lexicon, Sloleks

1 Uvod

Kot navajajo številni obstoječi slovenski jezikovni priročniki in jezikoslovne raziskave, je naglasno mesto v slovenščini večinoma nepredvidljivo (Toporišič idr. 2001: 105; Toporišič 2004: 66–67; Tivadar 2004) – pri tem obstajajo določene izjeme, tj. enote, ki so vedno (ali skoraj vedno) naglašene. V jezikovnih virih, kot sta *Slovenski pravopis* 2001 in *Slovenska slovnica*, so stalno naglašeni enotam sicer posvečeni ločeni odseki (več o tem v razdelku 2), v katerih so navedena tudi obrazila, ki so vedno naglašena, a ti seznami navajajo le nekaj primerov obrazil ter ponazoritvenih leksemov (s potencialnimi opombami, da je obrazilo redko). Pomanjkanje empiričnega gradiva in izčrpnih seznamov v obstoječih priročnikih je motivacija za raziskavo, ki jo predstavljamo v prispevku.

Za pripravo jezikovnega vira, ki bi natančneje popisoval naglašene enote v slovenščini (zlasti v strojno berljivi obliki), je argumentov več. V

projektu *Razvoj slovenščine v digitalnem okolju* (RSDO)¹ je bil posodobljen *Slovenski oblikoslovni leksikon Sloleks* (Čibej idr. 2022). V posodobitvi z različice 2.0 na 3.0 je bilo dodanih približno 265.000 novih leksemov, pri katerih so bile strojno generirane pregibne in naglašene oblike. Trenutno uporabljeni strojni naglaševalnik temelji na tehnologiji nevronske mreže (Krsnik 2017) in je bil naučen ter evalviran na manjši učni množici (Šef in Gams 2004),² ob posodobitvi pa bi ga bilo smiselno nadgraditi z dodatno stopnjo preverjanja stalno naglašanih enot (bodi si s hevrističnim preprocesiranjem ali postprocesiranjem za preverjanje sumljivih rezultatov – če npr. v naglaševalnik pošljemo besedo *čofotalniški* in kot rezultat prejmemo npr. naglašeno obliko *čofótalniški*, lahko z navzkrižno primerjavo s seznamom stalno naglašanih enot hitro zaznamo, da gre za verjetno napako, saj se beseda konča na *-alniški*, ki je stalno naglašen kot *-álniški*), kar bi izboljšalo postopek tudi za nadaljnje posodobitve leksikona. Tudi sicer je govor vse pogostejše obravnavan pri razvoju jezikovnih tehnologij (podatke o naglašanih oblikah potrebujejo tudi npr. sintetizatorji govora – gl. npr. Šef 2006), kar nakazuje tudi v času pisanja članka še potekajoči projekt *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE),³ ki med cilji navaja učinkovitejši razvoj odprto dostopnih govornih virov, ki so pomembna podlaga za razvoj govornih tehnologij.

Obstoječi sezname stalno naglašanih enot pogosto izpustijo tudi implicitne informacije, za katere predvidevajo, da so maternim govorcem razvidne – navedene so npr. enote *-álnik*, *-íst* in *-ilna* (*računálnik*, *absolutíst*, *gostílna*), ne pa tudi vse enote, ki so iz njih izpeljane (npr. *računálnik* → *računálniški* → *računálniškost*; *absolutízem* → *absolutíst* → *absolutístičen* → *absolutístičnost*; *gostílna* → *gostílničar* → *gostílničarski*). Ti podatki morajo biti eksplicitni tako za strojno obdelavo kot tudi za človeške uporabnike, zlasti za tiste, ki se učijo slovenščino kot tuji jezik, saj jim npr. ohranjanje mesta naglasa preko besedotvornih povezav ni nujno intuitivno (gl. npr. Stergar 2022).

¹ Spletna stran projekta *Razvoj slovenščine v digitalnem okolju* (RSDO): <https://rsdo.slovenscina.eu/>.

² Naglaševalnik na podlagi nevronske mreže dosega točnost v višini 87,62 %.

³ Spletna stran projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE): <https://mezzanine.um.si/>.

Obenem se je slovensko jezikoslovje do zdaj osredotočalo le na dihotomijo med stalno (ali skoraj vedno) naglašeni enotami ter vedno nenaglašeni enotami, ni pa še na voljo podrobnejših podatkov o tem, kako pogosto so naglašene ostale naglašene enote (npr. koliko je vseh leksemov z določeno naglašeno enoto oz. katere so izjeme), katerim besednovrstnim kategorijam pripadajo (in katere so njihove oblikoskladenjske značilnosti) ipd.

Cilj prispevka je torej izgradnja odprto dostopnega strojno berljivega seznama naglašeni enot v slovenščini, za kar uporabimo ročno pregledani del *Slovenskega oblikoslovnega leksikona Sloleks 3.0* (Čibej idr. 2022). V prispevku najprej preverimo, kako so stalno naglašene enote obravnavane v obstoječih jezikovnih virih (razdelek 2). V razdelku 3 predstavimo Sloleks kot vir podatkov, ki jih uporabimo za gradnjo seznama, postopek gradnje in strukturo končnih seznamov. V razdelku 4 opravimo analizo po besednih vrstah in pridobljene rezultate primerjamo z drugimi jezikovnimi priročniki, v razdelku 5 pa prispevek sklenemo s koraki za prihodnje delo.

2 Naglašene enote v obstoječih virih

V *Slovenskem pravopisu* 2001 so stalno naglašene enote omenjene v razdelkih 896–900 (Toporišič idr. 2001: 105) kot del pravil, ki v nekaterih primerih pomagajo določati le deloma predvidljivo naglasno mesto v slovenščini. Kot ponazoritev je eksplicitno navedena le peščica stalno naglašeni priponskih obrazil za različne besedne vrste (*-álec*, *-áč*, *-ánd* za samostalnike, *-álen*, *-ér*, *-ovít* za pridevnike, *-izíratí/-írati*, *-ávati*, *-otátí/-etátí* za glagole).

Izčrpnejši sezname stalno naglašeni besedotvorni in oblikotvorni morfemov⁴ po besednih vrstah so na voljo v *Slovenski slovnici* (Toporišič 2004: 66–67; za podobne sezname gl. tudi Toporišič 2006: 137–139), kjer pri določenih enotah najdemo tudi opombe o pogostosti (zelo redko), izjemah (npr. *-áti*, *-ím*, izjema: *slišati*, *slišim*) in o morebitnih konotacijah (*-ànt*, *zabušànt*; slabšalno). Obrazila so še nekoliko podrobneje našeta v poglavju o besedotvorju (Toporišič 2004: 165), a seznama nista povsem

⁴ Pri pripravi prve različice našega seznama se osredotočamo na naglašene enote na nivoju osnovne oblike posameznega leksema, ne na naglašene enote v vseh pregibnih oblikah.

usklajena – v poglavju o stalno naglašanih enotah npr. ni navedena enota *-áš* (*glagoljáš*), ki pa je navedena v poglavju o besedotvorju, a tam ni eksplicitno navedena kot stalno naglašena.

Omeniti je treba še nastajajoči *Pravopis* 8.0, ki pa v poglavju VII/1 (Slovnčni oris za pravopis – Glasoslovni oris) v razdelku o naglasnem mestu navaja predvsem to, da v slovenščini ni preprostega pravila o naglasnem mestu in da se pri naglaševanju ravnamo po normativnih slovarjih ali slovnica. Pri tem stalno naglašene enote niso omenjene.

3 Priprava podatkov in metodologija

Pri pripravi seznama naglašanih enot smo izhajali iz *Slovenskega oblikoslovnega leksikona Sloleks* (natančneje iz različice 3.0; Čibej idr. 2022), ki je največja odprto dostopna podatkovna množica s strojno berljivimi podatki o slovenskih leksemih. Vsak leksem v Sloleksu vsebuje tudi podatke o pregibnih oblikah in njihovih oblikoskladenjskih oznakah po sistemu Multext East v6,⁵ ki je v uporabi tudi za označevanje slovenskih korpusov (npr. Gigafida 2.0; Krek idr. 2020). Vsaka od pregibnih oblik (npr. *mize*) je opremljena tudi z eno ali več naglašeni oblikami (*míze*) – omeniti je treba, da gre pri tem za jakostno naglaševanje, ne tonemsko – in ustrezajočimi fonetičnimi transkripcijami (IPA: /^hmi:zɛ/, X-SAMPA: /^hmi:zE/). Različica 3.0 vsebuje 365.340 leksemov, od katerih je bilo približno 100.000 leksemov ročno pregledanih, popravljena pa so bila tudi naglasna znamenja v njihovih naglašanih oblikah.⁶ Pri izvozu seznama smo upoštevali le ročno preverjene lekseme.

Veliki večini obravnavanih leksemov (približno 98 %)⁷ so pripisane tudi kode oblikoslovnih vzorcev, ki ponazarjajo strojno berljive pregibne paradigme (za podrobnejša pojasnila gl. Arhar Holdt 2021; Arhar Holdt in

⁵ Oblikoskladenjske specifikacije za slovenščino Multext East v6: <https://nl.ijs.si/ME/V6/msd/html/msd-sl.html>.

⁶ Od skupno 2.792.003 naglašanih oblik je bilo v okviru projekta RSDO kot ustreznih potrjenih približno 86 % naglašanih oblik, 13 % pa je bilo treba ročno popraviti (podrobnejša evalvacija je objavljena v Čibej 2023).

⁷ Pri nekaterih leksemih so pregibne oblike še nerazrešene, npr. pri nekaterih pridevniki je treba korpusno potrditi stopnjevane oblike; tem leksemom v trenutni različici leksikona še nismo pripisovali oblikoslovnega vzorca.

Čibej 2018). Ker so nekatere stalno naglašene enote v *Slovenski slovnici* navedene skupaj s pregibnimi oblikami (npr. *-ováti/-újem*), smo pri izvozu upoštevali tudi oblikoslovne vzorce.

Izvozni algoritem poteka na naslednji način: pri vsakem leksemu najprej preverimo, koliko različnih naglašanih oblik je v leksikonu na voljo za osnovno obliko leksema (npr. nedoločnik pri glagolu *kupiti* ima na voljo dve naglašeni obliki: *kúpiti* in *kupíti*).⁸ Nato pri vsaki od naglašanih oblik v smeri od konca proti začetku naglašene oblike preverimo, na katerem grafemu se pojavi naglasno znamenje, in zabeležimo besedni del od zadnjega grafema z naglasnim znamenjem naprej (v primeru leksema *kupiti* torej *-úpiti* in *-íti*). Na ta način izdelamo frekvenčni seznam naglašanih končnih besednih delov⁹ (skupaj z oblikoskladenjskimi značilnostmi in oblikoslovnimi vzorci leksemov, v katerih se naglašeni besedni del pojavlja). Naglašeni besedni del *-ánje* se npr. pojavi v 1.109 naglašanih osnovnih oblikah pri občnih samostalnikih srednjega spola z oblikoslovnim vzorcem *Ss1.I.o.* Iz približno 100.000 obravnavanih leksemov smo na ta način izluščili 24.188 kombinacij končnega naglašenelega dela in pripadajočega oblikoslovnega vzorca; med temi se jih 66 % pojavi le enkrat, v vsaj 10 naglašanih oblikah pa se jih pojavi le 1.390 (5,7 %).

V naslednjem koraku za vsakega od izluščenih naglašanih delov najprej izdelamo njegovo nenaglašeno ustreznico (npr. *-ánje* → *-anje*; *-ítev* → *-itev*; *-ánski* → *-anski*) ter v vseh v analizo vključenih leksemih preverimo, koliko od njihovih nenaglašanih osnovnih pregibnih oblik se konča na nenaglašeno različico. Podatek nato uporabimo za izračun deleža naglašeniosti, ki nakazuje, koliko odstotkov leksemov, ki se končajo na nenaglašeni končni besedni del (npr. *-itev*), se hkrati konča tudi na naglašeni končni besedni del (npr. *-ítev*).¹⁰ Obenem zabeležimo primere, ki so naglašeni po pričakovanjih, in izjeme, ki od tega odstopajo.

⁸ Pri leksemih, ki imajo v Sloleksu navedeno možnost večnaglasnega izgovora, smo upoštevali tako večnaglasniške kot enonaglasniške naglašene oblike.

⁹ V prispevku uporabljamo izraz *končni besedni del*, saj je postopek luščenja zelo robusten in formaliziran. Pri luščenju torej ne gre nujno za končaje, obrazila ali podobno.

¹⁰ Pri deležu naglašeniosti upoštevamo tako naglasno mesto kot naglasno znamenje. 100-odstotni delež naglašeniosti pri besednem delu *-álnik* pomeni, da se prav vse naglašene oblike vseh pregibnih oblik, ki se končajo na *-alnik*, končajo z naglašenim besednim delom *-álnik*. Delež naglašeniosti pri besednem delu *-ént* je manjši (gl. razdelek 4.1), saj se nekatere naglašene oblike v tem primeru končajo na *-ènt*.

MTE-6	Vzorec	Naglašeni del	f _A	f _B	Delež naglašenosti	Primeri	Izjeme
Sos	Ss1.1.o	ánje	1.109	4.134	0,27	posploševanje, opsovaje, mečkaje ...	odjédanje, poležávaje, toníranje, ...
Ggd	G1.2.d	íti	1.037	2.220	0,47	naplastíti, zaposlíti, preobložíti, ...	ustolícíti, izprázniti, razžáliti, ...
Sos	Ss1.1.o	íranje	960	960	1,00	koordinírhanje, paradírhanje, pigmentírhanje, ...	/
Ggv	G1.1.v	írati	872	872	1,00	masakrírati, hipnotizírati, kalkírati, ...	/
Soz	Sz3.1.o	ítev	796	796	1,00	zatesnítev, zadolžitev, ozemljítev, ...	/

Tabela 1: Prvih pet vrstic seznama izvoženih naglašanih besednih delov iz *Slovenskega oblikoslovnega leksikona Sloleks*.

Nekaj primerov iz surovega seznama, na podlagi katerega smo opravili podrobnejšo analizo (več o tem v razdelku 4), je navedenih v tabeli 1. V stolpcih so navedeni naslednji podatki: osnovne oblikoskladenjske značilnosti leksema po sistemu Multext East v6 (npr. *Sos* – samostalnik, občni, srednji spol); koda oblikoslovnega vzorca; naglašeni končni besedni del; število naglašanih osnovnih oblik, ki se končajo na naglašeni končni besedni del (f_A); število nenaglašanih osnovnih oblik, ki se končajo na nenaglašeno različico naglašene končnega besednega dela (f_B); delež naglašenosti za posamezni naglašeni končni besedni del znotraj leksemov z enakim oblikoslovnim vzorcem; do 10 naključno izbranih primerov, ki vsebujejo naglašeni končni besedni del; in do 10 naključno izbranih izjem, ki se sicer končajo na nenaglašeno različico naglašene končnega besednega dela, ne vsebujejo pa samega naglašene končnega besednega dela.

4 Analiza

Pridobljeni surovi seznam naglašanih končnih besednih delov, razvrščenih po pogostosti, smo ročno analizirali po skupinah glede na osnovne oblikoskladenjske značilnosti, in izdelali nov, urejen seznam – zanimalo so nas predvsem stalno (oz. skoraj vedno) naglašene enote, zato smo se pri analizi omejili na naglašene končne besedne dele z deležem naglašenosti 80 odstotkov ali več – mejo smo določili tako, da je bila primerljiva s točnostjo trenutnega naglaševalnika in je hkrati ustrezno omejila količino gradiva za analizo. Na seznam smo vključili tudi nekaj izjem z nižjim deležem naglašenosti, saj kažejo na potencialne nekonsistentnosti v trenutni različici leksikona in jih je vredno analizirati podrobneje (npr. *-ák* pri pridevniki: *kroták*, *sladák* – več o tem v razdelku 4.2).

Relevantne naglašene končne besedne dele smo še dodatno razvrstili v štiri kategorije. Kategorija 1 vsebuje končne besedne dele, ki vedno ali v veliki večini primerov sovpadajo z besedotvornimi obrazili (oz. s kombinacijami besedotvornih obrazil; npr. *-álec*, *-ilničar* za samostalnike moškega spola). Kategorija 2 vsebuje naglašene končne besedne dele, ki ne sovpadajo z besedotvornimi obrazili, a so glede na podatke v leksikonu vedno ali zelo pogosto naglašeni (npr. *-óžec* pri samostalnikih moškega spola: *škrgonóžec*, *brezbóžec*, *trdokóžec*). Kategorija 3 vsebuje dele, ki so sicer vedno ali pogosto naglašeni, ne sovpadajo pa vselej z besedotvornimi obrazili (npr.

-áj pri samostalniki moškega spola, kjer je -áj kot besedotvorno obrabljeno pri vezáj, ne pa pri bonsáj). Kategorija 4 vsebuje izjeme in posebne primere, kot sta npr. -éntov/-éntov pri svojilnih pridevniki, ki sta sicer vedno naglašena, a je zaradi variantnosti njun delež naglašenosti nižji od 80-odstotnega praga.

V nadaljevanju zaradi prostorskih omejitev predstavljamo le poglavitne izsledke analize za glavne skupine leksemov. Celotni sezname so za podrobnejše analize na voljo na repozitoriju CLARIN.SI (več o tem v razdelku 5).

4.1 Občnoimenski samostalniki

Pri občnoimenskih samostalnikih smo zabeležili skupno 514 naglašanih končnih besednih delov, ki so naglašeni v vsaj 80 % primerov. Razporeditev po spolu in pripisanih kategorijah prikazuje tabela 2. V nadaljevanju zaradi prostorskih omejitev navajamo predvsem primere iz kategorije 1, ki je zastopana tudi v obstoječih jezikovnih virih.

Spol	Kategorija 1	Kategorija 2	Kategorija 3	Kategorija 4
Moški	41	93	27	1
Ženski	53	148	33	4
Srednji	27	69	18	/
Skupaj	121	310	78	5

Tabela 2: Razporeditev naglašanih končnih besednih delov pri občnoimenskih samostalnikih po spolu in kategoriji.

Pri moškem spolu je bilo v *Slovenski slovnici* in *Slovenskem pravopisu* 2001 med stalno naglašeni enotami omenjenih le 31 naglašanih končnih besednih delov – npr. -íst (*formalist*), -ízem (*parlamentarizem*). Med besednimi deli, ki smo jih izluščili, so stalno naglašeni (s 100 % deležem naglašenosti) še npr. -átor (*respirátor*), -áš (*revolveráš*), -órij (*krematórij*), -árij (*akvárij*) in -ítor (*investítor*). Številni so sestavljeni in izvirajo iz stalno naglašanih enot pri samostalnikih ženskega ali srednjega spola oz. pri drugih besednih vrstah: -íranec (*aretíranec*), -ítek (*ulítek*), -álkar (*ožigálkar*), -álničar (*zavaroválničar*), -iščnik (*semeníščnik*). Pri nekaterih naletimo na nekaj izjem, npr. -ívec (*opravljívec*, 99 %, izjema: *nadiúšljívec*), -ínar (*kovínar*, 94 %; izjemi: *veterínár*, *plinár*).

Za samostalnike ženskega spola referenčna vira omenjata 17 naglašanih končnih besednih delov. Kot stalno naglašene smo izluščili še npr. *-áva* (*izdeláva*), *-áza* (*špijonáza*), *-ánca* (*resonánca*), *-étev* (*klétev*) ter veliko število sestavljenih enot, ki izhajajo iz stalno naglašanih enot drugih besednih vrst (oz. iz samostalnikov drugega spola): *-ístka* (*sindikalistka*), *-álnost* (*ukazoválnost*), *-íranost* (*argumentíranost*), *-ílka* (*cedílka*), *-ívka* (*volívka*), *-ítost* (*korenítost*) ipd. Najdemo tudi primere enot z redkimi izjemami, ki so naglašene drugače, npr. *-ívost* (*vsiljívost*, 99,5 %; izjema: *nadúšljivost*), *-álka* (*obiskováka*, 99 %; izjeme: *neándertalka*, *símental-ka*), *-ívnost* (*ofenzívnost*, 97 %; *póзитivnost*, *skrivnóst*).

Pri srednjem spolu referenčna vira skupaj omenjata le šest stalno naglašanih enot (npr. *-álo*, *-ílo*, *-íšče*); med izluščenimi najdemo še sestavljene: *-íranje* (*koordiníranje*), *-ítje* (*izlítje*), *-áštvo* (*stiskáštv*), *-ílce* (*poročílce*), *-álce* (*bodálc*), *-álstvo* (*brálstvo*), *-íteljstvo* (*ljubíteljstvo*), *-úštvo* (*potepúštvo*), *-álníštvo* (*zavaroválníštvo*) ipd.

Omeniti je sicer treba, da vseh naglašanih končnih besednih delov, ki jih navajata *Slovenska slovnica* in *Slovenski pravopis* 2001, ni na našem končnem seznamu, saj bodisi v Sloleksu niso dovolj pogosti bodisi njihov delež naglašenosti ne presega 80 %. Enote, ki jih na končnem seznamu ni (pri teh navajamo ponazoritvene lekseme iz *Slovenske slovnice*), so za moške samostalnike npr. *-ál* (*lokál*), *-èr* (*cepèr*), *-ír* (*hudír*), *-úr* (*mehúr*), *-ón* (*potegón*), *-ávt* (*vohávt*); za ženske samostalnike npr. *-íra* (*sekíra*), *-ára* (*punčára*), *-éla* (*brucéla*), za srednje *-évje* (*pečévje*).

Ker je naša luščilna metoda zasnovana zelo robustno in naglašene končne besedne dele išče od zadnjega grafema z naglasnim znamenjem naprej, nekaterih enot ne izlušči na enak način, kot so navedene v *Slovenski slovnici*: tam je npr. za občne samostalnike ženskega spola navedenih 11 sorodnih enot, ki se ne začnejo z grafemom z naglasnim znamenjem: *-nína* (*povrtnína*), *-arína* (*brodarína*), *-enína* (*pletenína*), *-ečina* (*bolečina*), *-lína* (*oteklína*), *-evína* (*pajčevína*), *-ovína* (*steklovína*), *-tvína* (*kletvína*), *-avína* (*okopavína*), *-otína* (*dragotína*) in *-anína* (*tkanína*). V našem luščenju so vsi podobni primeri navedeni pod naglasnim končnim besednim delom *-ína*, ki ima delež naglašenosti 87 %, uvrščen pa je v kategorijo 3 (pokriva torej mešane primere, besedotvorne – *enajstína* – in naključne – *margarína*). Za natančnejšo preverbo enot

iz *Slovenske slovnice* bi bilo potrebno ločeno luščenje, kar smo pustili za prihodnje delo.

Med izjemami (kategorija 4) velja omeniti besedne dele *-ént/-ènt* in *-éntstvo/-èntstvo*. Ker gre za variante, dosegajo delež naglašenosti le okrog 50 %. Naglašene oblike v leksikonu, ki vsebujejo te dele, je treba še podrobneje preveriti in posodobiti.

Omeniti je treba tudi, da med ročno pregledanimi leksemi v Sloleksu 3.0 primanjkuje ženskih oblik, zato je zlasti pri samostalniki ženskega spola treba upoštevati, da seznam končnih besednih delov ni izčrpen, saj je zanje v trenutni različici leksikona še premalo primerov (npr. *doktorándka*, *magistrándka*). Pričakujemo, da bo s širjenjem leksikona odpravljena tudi ta pomanjkljivost.

4.2 Pridevniki in prislovi

Pri pridevnikih je bilo v *Slovenski slovnici* in *Slovenskem pravopisu* 2001 omenjenih 14 naglašanih enot z našega seznama, analiza pa je razkrila 231 naglašanih končnih besednih delov z deležem naglašenosti vsaj 80 %. Razporeditev po vrsti pridevnika in različnih kategorijah je prikazana v tabeli 3.

Vrsta	Kategorija 1	Kategorija 2	Kategorija 3	Kategorija 4
Splošni	24	97	24	3
Deležniški	4	30	1	/
Svojilni	21	18	8	1
Skupaj	49	145	33	4

Tabela 3: Razporeditev naglašanih končnih besednih delov po vrsti pridevnika in kategoriji.

Za splošne pridevnike smo poleg sedmih v virih že zabeleženih enot (npr. *-álen*, *-ív*, *-ílen*) izluščili še več sestavljenih, npr. *-ístičen* (*strukturalističen*), *-álski* (*prekupčeválski*), *-íran* (*talentíran*, *nèmotivíran*¹¹), *-ílski* (*morílski*), *-ántski* (*kapitulántski*), *-átorski* (*provokátorski*), *-iščen* (*zabaviščen*), *-íteljski* (*upravíteljski*), *-ílنيški* (*pomnílنيški*), *-álniški* (*igrálniški*). Na tej

¹¹ Zanikani deležniški pridevniki so po sistemu Multext East v6 obravnavani kot splošni.

točki se luščenje izkaže koristno tudi kot opozorilo na morebitne napake v leksikonu – pri naglašenem končnem besednem delu *-íran* je npr. delež naglašenosti le 98-odstoten, vzrok za to pa je neustrezno naglašena oblika med izjemami (*nèmaterialíziran*), ki bo popravljena v prihodnjih različicah leksikona.

Na seznam smo v kategorijo 4 uvrstili tudi naglašena končna besedna dela *-ák* (*kroták, sladák*) in *-án* (*dolžán*), ki sicer po naši metodi ne dosežata 80-odstotnega deleža naglašenosti, a je težava predvsem v tem, da so leme leksemov v leksikonu vedno pripisane z alternativama *-ek* (*krótek, sládek*) in *-en* (*dólžen*). Deleža naglašenosti zato po našem principu ni mogoče izračunati, saj v leksikonu formalno ne obstaja noben pridevniški leksem na *-ák/-án*.

Pri deležniških pridevniki smo izluščili štiri stalno naglašene enote, ki so že omenjene v referenčnih virih: *-íran* (*distribuíran*), *-óč* (*oklepajóč*), *-ét* (*prizadét*), *-ít* (*našít*); zanimivi pa so tudi besedni deli iz kategorije 2, med katerimi so stalno naglašeni npr. *-írt* (*odstírt, podírt*), *-ínjen* (*zedínjen, prekínjen, utekočinjen*), *-ézan* (*preplézan, obvézan, odrézan*); nekateri pa imajo vsaj zelo visoke deleže naglašenosti: *-ísan* (*porísan, zakísan*, 97 %; izjema: *razpísán*), *-ávljen* (*nadišávljen, odjávljen*, 89 %; izjema: *prekrvavljèn*).

Pri svojilnih pridevniki je primerov iz kategorije 1 več, a po pričakovanju vsi izhajajo iz stalno naglašanih enot samostalnikov, npr. *-ístov* (*akvarelistov*), *-áčev* (*kováčev*, 99 %; izjema: *Bílačev*), *-íteljev* (*vzgojíteljev*), *-átorjev* (*administrátorjev*), *-álčev* (*tekmováčev*; izjema: *neándertalčev*), *-íčkov* (*fantíčkov*), *-ákin* (*plesákin*), *-áteljičin* (*pisáteljičin*).

Podobno kot pri samostalnikih (gl. razdelek 4.1) se tudi tu pojavi izjema *-éntov/-éntov* z deležem naglašenosti 50 %, ki v prihodnje zahteva podrobnejšo obravnavo.

Pri splošnih prislovi (tabela 4) smo izluščili podobne naglašene končne besedne dele kot pri splošnih pridevniki: *-álno* (*emocionálno*), *-ístično* (*dokumentarístično*) ipd. Pri deležniških prislovi poleg enot, ki so že omenjene v referenčnih virih, velja omeniti še stalno naglašene, a po pogostosti redke *-ívši* (*stopívši*), *-énši* (*zacénši*), *-ávši* (*spoznávši*) in *-évši* (*vštévši*).

Vrsta	Kategorija 1	Kategorija 2	Kategorija 3	Kategorija 4
Splošni	26	55	11	/
Deležniški	7	/	/	/
Skupaj	33	55	11	/

Tabela 4: Razporeditev naglašanih končnih besednih delov po vrsti prislova in kategoriji.

4.3 Glagoli

Posebej smo luščili naglašene končne besedne dele za dovršne, nedovršne in dvovidske glagole (tabela 5); po sistemu Multext-East v6 so to namreč oblikoskladenjske značilnosti, ki ločujejo različne vrste leksemov.

Vrsta	Kategorija 1	Kategorija 2	Kategorija 3	Kategorija 4
Dovršni	33	36	2	-
Nedovršni	10	37	1	-
Dvovidski	10	6	1	-
Skupaj	53	79	4	-

Tabela 5: Razporeditev naglašanih končnih besednih delov po glagolskem vidu in kategoriji.

Poleg v referenčnih virih že zabeleženih *-ávati* (*izžvižgávati*), *-írati* (*zabetonírati*), *-éti* (v omejenem številu oblikoslovnih vzorcev, npr. *odletéti* – *odletím*) ipd. je bilo izluščenih več kombinacij naglašene končnega dela *-áti* in posameznih oblikoslovnih vzorcev: *-áti* v kombinaciji z vzorcem *G2.2.v* denimo ustreza stalni naglašeni enoti *-ováti* (*predrugáčeváti*, *ozelenjeváti*, *osrčeváti*). Tako kot pri samostalniškem naglašenem končnem besednem delu *-ína* (gl. razdelek 4.1) je tudi pri glagolih zato potrebno še dodatno luščenje naglašanih besednih delov, ki se ne začnejo z grafemom z naglasnim znamenjem. Te bomo podrobneje analizirali ob prihodnjih posodobitvah seznama. Omenimo lahko še *-úti* (*vsúti*, *izrúti*, *obúti*), ki v referenčnih virih ni naveden.

5 Zaključek in prihodnje delo

V prispevku smo predstavili izdelavo seznamov slovenskih naglašanih končnih besednih delov na podlagi *Slovenskega oblikoslovnega leksikona Sloleks* in poglavitne izsledke analize izluščenih naglašanih besednih delov. Seznama – tako surovi izvoz kot urejeni in kategorizirani nabor – sta pod nazivom SNES 1.0 (Slovenske naglašene enote iz Sloleksa) naložena na repozitoriju CLARIN.SI pod odprto licenco Creative Commons BY-SA 4.0 (Čibej 2024). Gre za vir, ki ga je mogoče dinamično spreminjati in posodabljati vzporedno z dopolnjevanjem in popravljanjem leksikona, koristen pa je lahko tako za snovalce in snovalke slovenskih jezikovnih priročnikov kot tudi za potrebe strojnega naglaševanja oz. preverjanja rezultatov strojnega naglaševanja.

Ker smo pri gradnji trenutne različice leksikona upoštevali le stalno naglašene enote v osnovnih oblikah leksemov, bi bilo smiselno v prihodnje SNES nadgraditi. Kot prvo, s podobno metodologijo je mogoče izluščiti tudi seznam naglašanih enot za posamezne naglašene pregibne oblike (npr. sedanjik ednine pri glagolih na *-ováti: -újem*). Kot drugo bi bilo koristno izvoziti tudi seznam nenaglašanih končnih besednih delov, tako da npr. iz nenaglašanih pregibnih oblik izvozimo končne besedne dele ter preverimo, kateri se nikoli ne pojavljajo med naglašeni.

Kot smo že omenili v razdelku 4, bi bilo treba podrobneje preveriti enote, ki se ne začnejo z grafemom z naglasnim znamenjem (npr. *-ováti, -nína*) ter še podrobneje ločevati med besedotvornimi obrazili in naključnimi naglašeni končnimi besednimi deli (*enajstína* vs. *margarína*). Ta naloga je povezana tudi z morfološko segmentacijo materiala, ki je vključen v Sloleks, saj bi bilo na segmentiranih leksemit lažje upoštevati morfemske meje in ločevati med naključnimi naglašeni besednimi deli in morfemi.

Če na problematiko naglaševanja v slovenščini pogledamo širše in ne zgolj z vidika naglašanih enot, pa bi bilo v prihodnje smiselno opraviti še nekoliko bolj poglobljeno statistično analizo (npr. naglasna mesta po zlogih in po besednih vrstah), ki bi morda kljub dejstvu, da za naglaševanje v slovenščini ni enostavnega pravila, razkrila določene značilnosti, ki so v slovenščini vsaj do neke mere predvidljive (npr. ali naglasno mesto korelira s številom zlogov, z besedno vrsto ipd.).

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (J7-4642), raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411) in raziskovalne infrastrukture CLARIN.SI, ki jih financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS). Iskrena hvala tudi anonimnima recenzentoma za konstruktivne pripombe.

Viri

Čibej, J. (2024). *Lists of Slovene accentuated units SNES 1.0* [Nabor podatkov]. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/2003>

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Krsnik, L. in Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 3.0* [Nabor podatkov]. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1745>

Pravopis 8.0: Pravila novega slovenskega pravopisa za javno razpravo. (2024). <https://pravopis8.fran.si/>

Toporišič, J. (2004). *Slovenska slovnica* [4. prenovljena in razširjena izdaja]. Obzorja.

Toporišič, J. idr. (ur.). (2001). *Slovenski pravopis*. Založba ZRC, ZRC SAZU. www.fran.si

Literatura

Arhar Holdt, Š. (2021). Oblikoslovni vzorci za strojno procesiranje slovenščine. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 87–124). Založba Univerze v Ljubljani. <https://doi.org/10.4312/9789610605478>

Arhar Holdt, Š. in Čibej, J. (2018). Oblikoslovni vzorci v leksikonu Sloleks. *Slovenščina* 2.0, 6(2), 33–66. <https://doi.org/10.4312/slo2.0.2018.2.33-66>

Čibej, J. (2023). *Leksikon besednih oblik Sloleks. Poročilo projekta Razvoj slovenščine v digitalnem okolju. Aktivnost DS1.3*. Center za jezikovne vire in tehnologije, Univerza v Ljubljani. https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_Sloleks_v2.pdf

- 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: May 11-16, 2020, Marseille, France* (str. 3340–3345). ELRA - European Language Resources Association.
- Krsnik, L. (2017). *Napovedovanje naglasa slovenskih besed z metodami strojnega učenja* [Magistrsko delo]. Univerza v Ljubljani.
- Stergar, M. (2022). *Analiza naglasnega mesta branih besedil tujih študentov UP: zaključno delo* [Diplomsko delo]. Univerza na Primorskem.
- Šef, T. (2006). Avtomatsko naglaševanje nepoznanih besed pri sintezi slovenskega govora. *Elektrotehniški vestnik*, 73(2–3), 99–104.
- Šef, T. in Gams, M. (2004). Data mining for creating accentuation rules. *Applied Artificial Intelligence*, 18(5), 395–410.
- Tivadar, H. (2004). Fonetično-fonološke lastnosti samoglasnikov v sodobnem knjižnem jeziku. *Slavistična revija*, 52(1), 31–48.
- Toporišič, J. (2006). *Besedjeslovne razprave*. Inštitut za slovenski jezik Frana Ramovša.