Množičenje za slovar sodobnega slovenskega jezika

Darja Fišer, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Iztok Kosem, Špela Arhar Holdt, Damjan Popič in Tomaž Erjavec

Abstract
Crowdsourcing will play an important part in the compilation of a new monolingual dictionary of Slovenian as a method for filtering and processing automatically extracted corpus data, which will then serve as a basis for the preparation of final dictionary entries by lexicographers. The success of a crowdsourcing campaign depends on a number of factors, e.g. effective workflow, the funding available, the technological framework for crowdsourcing, the interests of the crowdsourcers, and the type and volume of data to be processed. Before starting a project, it is imperative to analyse its needs and plan for the implementation of crowdsourcing under different conditions in order to ensure the feasibility of the campaign and the usefulness of its results. In this paper, a crowdsourcing workflow for lexicographic projects is suggested and different scenarios are discussed for implementing crowdsourcing in accordance with the project funds available. In addition to an overview of the most popular crowdsourcing platforms already used in similar projects, a discussion is also presented on the criteria that were taken into account when selecting the most appropriate platform for the needs of a specific lexicographic project. To conclude, a number of examples are provided to illustrate some of the potential uses of crowdsourcing in various phases of dictionary construction.

Keywords: crowdsourcing, workflow, dictionary construction, crowdsourcing platforms

Ključne besede: množičenje, delotok, gradnja slovarja, platforme za množičenje

1 UVOD

V načrtu za gradnjo slovarja sodobnega slovenskega jezika ima množičenje pomembno vlogo pri obdelavi avtomatsko izluščenih podatkov in njihovi pripravi za nadaljnji leksikografski postopek. Ker je uspešnost množičenja v veliki meri odvisna od številnih zunanjih dejavnikov, kot so npr. razpoložljiva sredstva, motivacija množičnikov in obseg ter vrsta podatkov, ki jih je treba obdelati, je pomembno, da se že pred začetkom projekta predvidi, kako bi tovrstna obdelava podatkov potekala v različnih okoliščinah.

V prispevku zato najprej predstavljamo splošni predlog delotoka množičenja za leksikografske projekte, v katerem je postopek obdelave podatkov razdeljen na stopnje, ki jih je glede na zahteve in zmožnosti projekta mogoče prilagoditi. Nadaljujemo z opisom možnih scenarijev vključevanja množičenja v projekt, ki so prilagojeni različnemu trajanju in obsegu financiranja. Pri vsakem navedemo ključne vidike, ki jih je treba upoštevati, če želimo izvesti uspešno množičenjsko kampanjo, in posameznemu scenariju prilagojene množičenjskega delotoka.

Opravimo pregled značilnosti dobrih platform za množičenje in opišemo najbolj razširjene platforme, ki so bile že uporabljene v sorodnih projektih, ter kriterije, po katerih smo izbrali platformo za množičenje, ki najbolj ustreza zahtevam načrtovanega leksikografskega projekta. Nazadnje predstavimo še predhodno analizo potreb in prve predloge mikronalog, ki bi jih bilo mogoče uporabiti pri različnih slovaropisnih stopnjah od izdelave leksikona do preverjanja uporabniške izkušnje s slovarskim vmesnikom.

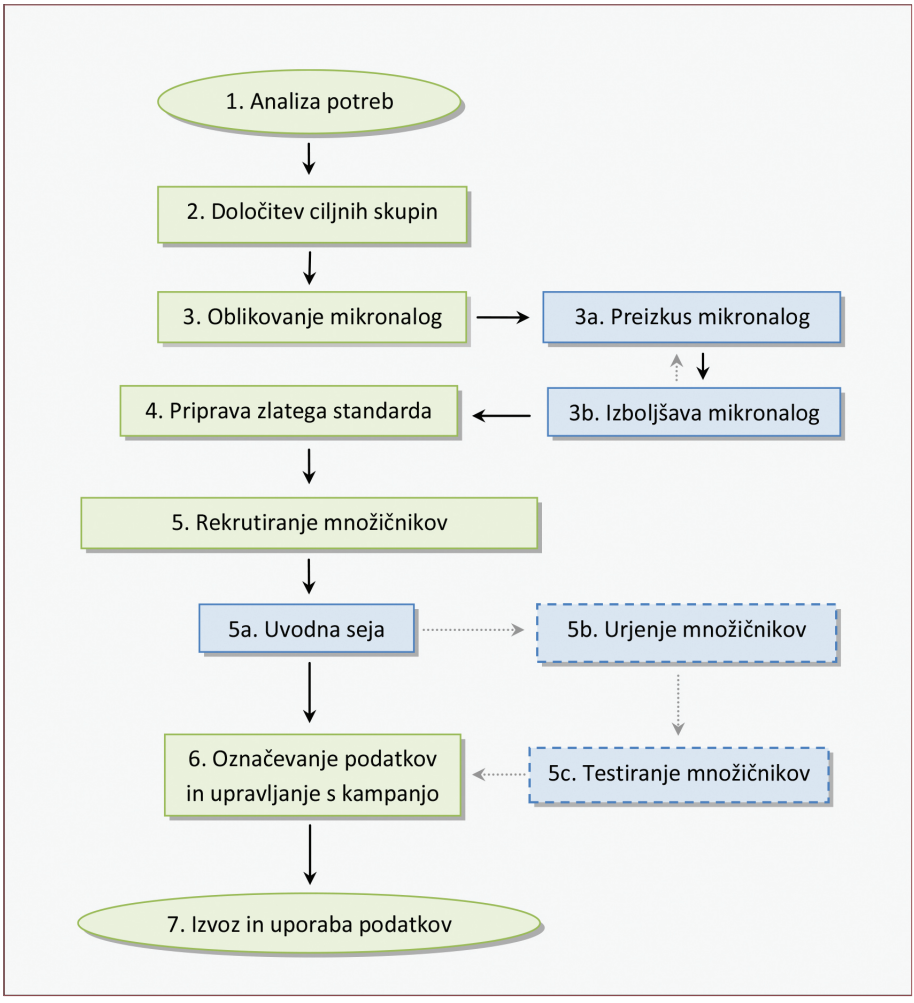
2 DELOTOK MNOŽIČENJA V LEKSIKOGRAFSKIH PROJEKTIH

V tem razdelku predstavljamo splošni predlog množičenjskega delotoka, ki smo ga zasnovali za uporabo pri različnih stopnjah korpusno podprtih leksikografskih projektov. Pristop je modularen in ga je mogoče prilagoditi glede na specifične zahteve projekta. Spreminjati je mogoče vrstni red posameznih stopenj in nekatere tudi izpustiti, a je pomembno, da kljub temu razmislimo o vseh vprašanjih, ki jih posamezne stopnje naslavljajo, saj množičenjska kampanja zahteva veliko truda, časa in sredstev, a se brez pozornega načrtovanja in upravljanja kaj lahko zgodi, da dobljeni rezultati na koncu sploh niso uporabni.

Pred začetkom vsake množičenjske kampanje je treba vnaprej in preudarno oceniti, koliko denarja, časa in delovne sile bo zahtevala. Kampanja namreč ni

smiselna, če zahteva več truda, časa in sredstev kot konvencionalne metode ročne obdelave podatkov. Pomembna prednost vključevanja množičenja v sam načrt gradnje slovarja pa je, da se bo tako začetni vložek v celostno pripravo ustreznega množičenjskega okolja kmalu povrnil, saj bo množičenje tako mogoče uporabiti v številnih fazah slovarskega projekta, mikronaloge zasnovati po enakih načelih, podatke pa se označi in obdela po enaki metodi in na isti platformi.

V nadaljevanju sledi opis posameznih stopenj delotoka.



Slika 1: Shema splošnega delotoka množičenja za leksikografske projekte. Z zeleno barvo so označene glavne faze, z modro pa podfaze delotoka. S polno črto so označene obvezne, s črtkano pa neobvezne faze delotoka.

1. Analiza potreb – Prvi korak vsake množičenjske kampanje je celovita analiza potreb. Treba je določiti cilje in uskladiti pričakovanja od kampanje, količino podatkov, ki jih je treba obdelati, namene, za katere bodo podatki uporabljeni ter komu, v kakšnem formatu in pod kakšnimi pogoji bodo rezultati na voljo. Pri slovarskih projektih, kjer se množičenje lahko uporabi v različnih fazah slovarskega projekta, je smiselno analizirati potrebe za vsakega od teh segmentov in nato delotok, platformo in časovnico množičenjskih kampanj zasnovati tako, da so vhodni podatki in vsa programska oprema s čim manjšo mero prilagajanja primerni za uporabo v vseh fazah projekta.

2. Določitev ciljnih skupin – Po analizi potreb je treba določiti zahtevani profil množičnikov, saj so naloge različno kompleksne in zahtevajo različno predznanje. Določene naloge lahko rešuje tudi splošna javnost brez specializiranega jezikoslovnega ali leksikografskega znanja, kompleksnejše naloge pa lahko učinkovito rešijo le ustrezno usposobljeni posamezniki (npr. študentje in diplomanti jezikovnih smeri ali celo leksikografi). Ključno je, da naloge damo v reševanje pravi skupini, saj je to predpogoj za kakovostne rezultate.

3. Oblikovanje, preizkušanje in urejanje mikronalog – Najpomembnejši in obenem najtežji korak vsake množičenjske kampanje je oblikovanje mikronalog. Vprašanja morajo biti jasna, kratka in enoznačna ter prilagojena predznanju ciljne skupine množičnikov. V vprašanjih za splošno javnost se tako na primer izogibamo strokovnim izrazom in kompleksnim strukturam, ki jih nadomeščamo s splošnimi oz. s praktičnimi primeri (npr. namesto »Kateri pomen najbolje ustreza rabi besede v zvezi, ki jo vsebuje zgled?« raje »Kaj pomeni podčrtana beseda v spodnjem stavku?«). Zelo pomembno je, da nalog ne zasnujemo tako, da bodo prinesle nezanesljive rezultate. Topogledno so še posebej problematična večdimenzionalna vprašanja, saj množičniki ne bodo vedeli, kako naj nanje odgovarjajo (npr. namesto »Je prikazana kolokacija ustrezna za vključitev v slovar?« nalogo raje razdelimo na dva dela: 1. »Je prikazana kolokacija pravilno izluščena iz korpusa?« in 2. (samo za pravilno izluščene) »Sodi prikazana kolokacija v učni slovar?«). Izdelane mikronaloge je treba pred množičenjem preizkusiti v pilotni raziskavi, da preverimo njihovo učinkovitost in določimo morebitne neskladnosti, nejasnosti in napake, ter vse identificirane slabosti pravočasno odpraviti. Če se katera od nalog izkaže kot preveč zapletena za profil množičnika, ki mu je bila določena, jo je treba bodisi prilagoditi bodisi dati v reševanje skupini z več predznanja.

4. Izdelava zlatega standarda – Določeno število mikronalog morajo označiti strokovnjaki, da lahko njihove odgovore uporabljamo za preverjanje zanesljivosti množičnikov. To zbirko referenčnih mikronalog imenujemo zlati standard, ki mora biti karseda reprezentativen, tako po velikosti kot tudi po težavnosti vsebovanih nalog glede na kompleksnost obravnavanega problema.

5. Rekrutiranje množičnikov – Po oblikovanju mikronalog in izdelavi zlatega standarda je treba najeti množičnike in jih seznaniti s postopkom označevanja. Pobudnik množičenja na začetku ponavadi organizira **uvodno sejo** (angl. *demo session*), ki včasih poteka v živo, največkrat pa v obliki predstavitve ali videoposnetka, dostopne na spletni strani projekta, ki množičnikom predstavi, kako poteka označevanje. Uvodni seji sledi **urjenje** (angl. *training session*), kjer množičniki rešujejo naloge v živo pod nadzorom strokovnjaka, ki svetuje in nudi dodatna pojasnila, ali pa na spletu, pri čemer ob vsaki rešeni nalogi prejmejo avtomatizirano povratno informacijo. Zadnji korak rekrutiranja je **testiranje** (angl. *testing session*), pri katerem množičniki naloge rešujejo brez pomoči, na podlagi točnosti njihovih rezultatov pa se odločimo, ali jih bomo najeli ali ne. Pri nizkoproračunskih projektih je urjenje in testiranje množičnikov pogosto združeno z glavnim delom kampanje, nezanesljive odgovore/množičnike pa se izloči naknadno.

6. Reševanje nalog in upravljanje s kampanjo – To je glavna faza vsake množičenjske kampanje, v kateri množičniki rešujejo mikronaloge. Pobudnik mora skrbno nadzirati potek kampanje in redno preverjati vmesne rezultate ter se odločati, ali so potrebni dodatni ukrepi, npr. ali je treba povečati število mikronalog, ali so množičniki dovolj motivirani, da naloge rešujejo redno, ali so rezultati v skladu s pričakovanji projekta ipd.

7. Izvoz in uporaba podatkov – V zadnjem koraku rezultate množičenja izvozimo v ustrezen format, ki omogoča naknadno filtriranje in nadaljnjo uporabo (npr. za učenje algoritmov ali za vključitev v slovar). Pomembno je, da platforma za množičenje omogoča izvoz podatkov tudi sredi kampanje, saj je sprotno preverjanje uporabnosti rezultatov ključno za učinkovito upravljanje s kampanjo.

3 VRSTE LEKSIKOGRAFSKIH PROJEKTOV

V tem razdelku predstavljamo potencialne scenarije vključevanja množičenja v različne tipe leksikografskih projektov. Kot smo že poudarili, je potek množičenjske kampanje v veliki meri odvisen od stopnje financiranja, saj se na podlagi tega pobudnik množičenja tudi odloča, v kakšnem obsegu in časovnem okviru bo množičenje izvajal, v katere faze leksikografskega dela ga bo vključeval, koliko tipov nalog bo za to razvil, kako kompleksen bo delotok množičenja, koliko množičnikov bo rekrutiral in na kakšen način jih bo motiviral. Več kot je na voljo financiranja, tem bolj specializirane aplikacije je zanj mogoče razviti, jih temeljito preizkusiti in jih z optimalnimi nastavitvami ponuditi v uporabo širokemu krogu ljudi. Po drugi strani pa je pri projektih s skromnimi finančnimi sredstvi potreben veliko večji poudarek na rekrutiranju in motivaciji množičnikov. Seveda je družbeno motivacijo mogoče (in priporočljivo) uporabiti tudi pri drugih scenarijih kot dodatno spodbudo za množičnike.

Slika 2: Možnosti uporabe množičenja v različnih vrstah leksikografskih projektov.

3.1 Specializirani projekti

Specializirani projekti s polnim financiranjem si lahko privoščijo razvoj ciljne aplikacije za množičenje. Pri tem je najbolj smiselno izkoristiti možnosti iger z namenom, ki z zabavnimi in tekmovalnimi elementi dolgoročno pritegnejo široko množico igralcev in izzovejo spontano rabo jezika, zaradi česar so se izkazale za uspešne v več sorodnih projektih (prim. Jurgens in Navigli 2014; Joubert in Lafourcade 2012; Chamberlain et al. 2008). Jurgens in Navigli (2014) celo ugotavljata, da igra z namenom Puzzle Racer, s katero igralci označujejo korpusne podatke, dosega enako kakovostno raven, kot če bi enake podatke označili strokovnjaki, obenem pa je postopek cenejši kar za 73 %, kot če bi podatke obdelali množičniki z reševanjem klasičnih nalog. Posebej razvita igra z namenom omogoča hitro zbiranje večjih količin podatkov, prilagoditi jo je mogoče za različne naprave in platforme ter vanjo vključiti različne naloge, namenjene za različne ciljne publike in faze specializiranega leksikografskega projekta.

3.2 Projekti s krovnim financiranjem

Veliko sodobnih leksikografskih projektov nima neposrednega financiranja, temveč se izvaja kot ena od neprimarnih dejavnosti v okviru širšega raziskovalnega

projekta oz. programa. V tem primeru je dejavnosti treba še posebej pazljivo izvajati tako, da rezultati tudi z močno omejenimi finančnimi in človeškimi viri izpolnijo zahteve tako krovnega kot leksikografskega projekta. V tem scenariju je smiselno z maksimalnim izkoristkom že obstoječih virov in tehnologij množičenjske kampanje zasnovati tako, da bodo poleg leksikografskega projekta neposredno uporabni tudi za druge namene v okviru krovnega oz. prihodnjih projektov. Ker v tem scenariju sredstev za razvoj ciljnih aplikacij najverjetneje ni, množičnikom naloge v reševanje ponudimo preko klasične platforme za množičenje. K sodelovanju skušamo pritegniti samostojne leksikografe, lektorje, prevajalce in ljubitelje jezika, ki za sodelovanje prejmejo mikroplačila, pri čemer so število nalog, obseg množičenih podatkov in količina najetih množičnikov sorazmerni z razpoložljivimi sredstvi. Po potrebi so iz delotoka izpuščene neključne faze (glej Sliko 1), kot je ciklično urejanje in izboljšave mikronalog ciklične ter urjenje in testiranje množičnikov.

3.3 Nizkoproračunski projekti

Večino sredstev, ki so za množičenje na voljo v nizkoproračunskih projektih, je smiselno vložiti v čim večjo avtomatizacijo priprave podatkov, delotok množičenja pa maksimalno poenostaviti. V tem scenariju tako npr. uporabimo privzete parametre za preverjanje zanesljivosti množičnika in pri sprejemanju končnih odločitev upoštevamo le večinski glas brez razsojanja strokovnjaka. Najprimernejši množičniki v tem scenariju so študentje jezikovnih smeri in ljubitelji, ki so za svoje delo nagrajeni z darilnimi boni, vstopnicami ali drugimi manjšimi materialnimi nagradami. Ta pristop se je že izkazal za izvedljivega (El-Haj et al. 2014; Fišer et al. 2014), a zahteva realna pričakovanja do množičnikov glede truda ter časa, ki so ga v kampanjo pripravljeni vložiti, zato jim ne dajemo zelo ambiciozno zastavljenih nalog. Od njih prav tako ne pričakujemo, da bodo v kratkem času opravili velike količine dela, kar je treba upoštevati že pri načrtovanju projekta, ki ga je treba zastaviti nekoliko bolj dolgoročno kot pri scenarijih s financiranjem.

3.4 Projekti brez financiranja

Kadar sredstev za množičenje ni, ga je mogoče izvesti na podoben način, kot to že uspešno počnejo številni kolaborativni leksikografski projekti, ki množičnike k sodelovanju pritegnejo izključno z nematerialnimi spodbudami (družbeno motivacijo). Poleg navdušenih posameznikov, ki jim je v veselje prispevati h gradnji novih jezikovnih virov za slovenščino, bi širšo javnost k reševanju nalog lahko

spodbujali tudi z nalogami, ki jih je zabavno reševati, ali s prirejanjem tekmovanj (npr. z uvedbo točkovnega sistema na izbrani platformi za množičenje, prim. Fišer et al. 2014), dijake, študente in mlade diplomante pa bi k sodelovanju lahko pritegnili tudi s priznanji za sodelovanje pri projektu, ki jih množičniki lahko izkoristijo za priznanje (ob)študijskih obveznosti ali navedejo kot referenco v življenjepisu.

Še bolj kot pri nizkoproračunskem scenariju je treba tak projekt zastaviti dolgoročno, brez časovnega pritiska in ne preveč ambiciozno. Množičnikom se v tem primeru reševanje ponudi samo najpreprostejše naloge, projekti pa morajo biti čimbolj relevantni za njihovo skupnost. Upoštevati je treba, da množičniki delo izvajajo iz lastnega navdušenja nad projektom, zato je še toliko bolj pomembno, da se redno vzdržuje stik in neguje dobre odnose z njimi ter gradi skupnost.

4 IZBIRA PLATFORME ZA MNOŽIČENJE

Platforma za množičenje je spletna aplikacija, na katero lahko pobudnik množičenja naloži projekt z mikronalogami, ki jih nato rešujejo najeti množičniki. V tem razdelku opišemo kriterije, na katere je treba biti pozoren pri izbiri platforme, in postopek, po katerem smo izbirali platformo, ki je predvidena za potrebe množičenja slovarja sodobnega slovenskega jezika.

4.1 Ključne značilnosti dobre platforme

Izbira ustrezne platforme je eden od prvih korakov pri množičenjski kampanji, pri odločitvi pa je treba upoštevati več kriterijev.

Format podatkov – Platforma mora omogočati nalaganje različnih tipov mikronalog in izvoz rezultatov množičenja v formatih, ki ustrezajo zahtevam projekta.

Vmesnik – Pomembno je, da platforma ponuja preprost, uporabniku prijazen vmesnik tako za za administratorja kampanje kot tudi za množičnike. Administratorju mora platforma omogočati, da oblikuje različne tipe različno kompleksnih nalog, med kampanjo spremlja statistiko zbiranja odgovorov in zanesljivost množičnikov, po potrebi nemoteče za množičnike razširi zlati standard ali posodobi bazo z mikronalogami in izvozi vmesne rezultate. Množičnikom mora biti omogočena preprosta registracija (npr. z računom Gmail, Twitter ali Facebook), varovanje osebnih podatkov in udobno delovno okolje, ki jim olajša delo in pozitivno vpliva na njihovo motivacijo.

Preverjanje kakovosti – Pomembno je, da platforma omogoča preverjanje kakovosti rezultatov množičenja s pomočjo različnih mehanizmov, kot so zlati standard, ujemanje med množičniki, doslednost množičnikov, večinski odgovor ipd. Prav tako je pomembno, da platforma omogoča spreminjanje parametrov za vključevanje vprašanj iz zlatega standarda, ponavljanje mikronalog pri različnih množičnikih, dovoljen čas za reševanje posamezne naloge ipd.

Finančni vidik – Če se za sodelovanje v kampanji predvideva mikroplačila, mora to izbrana platforma omogočati. Pri komercialnih platformah za množičenje, ki ponujajo gostovanje kampanj, je znesek, ki ga mora pobudnik množičenja nakazati, odvisen od velikosti in kompleksnosti kampanje. Večina nakazanega denarja se porabi za mikroplačila (njihovo višino ponavadi določi pobudnik sam), določen odstotek pa pobere upravljalec platforme.

Motivacijski mehanizmi – Prednost je, če ima platforma že vgrajene mehanizme za dodatno motivacijo množičnikov, kot npr. podeljevanje točk za pravilne odgovore, objava lestvic najuspešnejših množičnikov, avtomatska obvestila množičnikom, ko jih nekdo izrine z visokega mesta, in vabila množičnikom, ki v kampanji že dlje časa niso bili aktivni.

4.2 Pregled obstoječih platform za množičenje

Pri izbiri platforme za izdelavo slovarja slovenskega sodobnega jezika smo med oktobrom in novembrom 2014 opravili pregled okoli 150 platform za množičenje in izluščili tiste, ki omogočajo množičenje jezikovnih podatkov, in jih na kratko predstavljamo v nadaljevanju.

4.2.1 Plačljive platforme

Najbolj znana in tudi najpogosteje uporabljana platforma za množičenje je Amazon Mechanical Turk,[1] pri kateri so v administratorski vmesnik že vgrajena sredstva za preverjanje kakovosti, upravljanje z množičenjsko kampanjo in izplačevanje mikroplačil množičnikom. Na platformi je že registrirano veliko število množičnikov, a gre večinoma za materne govorce večjih jezikov.

Podobna primera sta Crowdflower[2] in Clickworker,[3] ki ponujata vrsto aplikacij za različna področja obdelave podatkov (npr. kategorizacija podatkov in analiza

1 https://www.mturk.com (dostop 8. 8. 2015).
2 http://www.crowdflower.com (dostop 8. 8. 2015).
3 http://www.clickworker.com/ (dostop 8. 8. 2015).

sentimenta). Mikronaloge je mogoče naložiti v jezikih CML, CSS ali Javascript, množičnike pa je mogoče filtrirati po starosti, predznanju in geografski lokaciji.

4.2.2 Odprtokodne platforme

Med odprtokodnimi platformami izstopa Crowdcrafting,[4] na kateri lahko množičniki prostovoljci z reševanjem nalog prispevajo k raziskovalnim projektom z različnih področij. Platforma temelji na tehnologiji PyBossa,[5] prosto dostopni programski opremi za ustvarjanje množičenjskih projektov, ki jo je mogoče namestiti na lokalni strežnik in je na voljo pod licenco Creative Commons BY-SA 4.0.

Prosto dostopno je tudi orodje sloWCrowd[6] (Tavčar et al. 2012), ki temelji na jeziku PHP/MySQL in je bilo razvito za namene čiščenja sloWNeta s pomočjo množičenja (Fišer et al. 2014).

4.3 Izbira platforme za izdelavo slovarja sodobnega slovenskega jezika

Po pregledu platform smo se odločili, da pri množičenju za izdelavo slovarja sodobnega slovenskega jezika uporabimo platformo PyBossa, in sicer iz naslednjih razlogov:

Prilagodljivost – PyBosso je za razliko od komercialnih platform mogoče namestiti na lokalni strežnik in vmesnik polno prilagoditi zahtevam in pogojem projekta.

Podprtost – PyBossa je kot odprtokodna platforma dobro podprta in se stalno razvija. Ker jo uporabljajo za mnoge projekte, so zanjo razvite tudi številne dodatne knjižnice, ki omogočajo več statističnih funkcij za spremljanje poteka množičenja, preverjanje kakovosti rezultatov ipd.

Finančna neodvisnost – V primeru nezadostnega financiranja projekta množičnikov ne bo mogoče plačati z mikroplačili, komercialne platforme pa ne omogočajo drugih oblik plačila (nagrad, vstopnic ipd.). Z uporabo odprtokodne platforme prihranimo tudi provizijo, ki jo pri izplačevanju mikroplačil zahtevajo komercialne platforme.

4 http://crowdcrafting.org/ (dostop 8. 8. 2015).
5 http://pybossa.com/ (dostop 8. 8. 2015).
6 http://nl.ijs.si/slowcrowd/ (dostop 8. 8. 2015).

Logistika – Pri nekaterih komercialnih platformah prihaja do logističnih zapletov, saj npr. na platformi Amazon Mechanical Turk pobudnik množičenja potrebuje odprt bančni račun v ZDA, če želi na platformi izvajati množičenjske kampanje. Platforma bi zahtevala predhodno registracijo in podatke tudi od vsakega slovenskega množičnika, kar je zelo neprikladno. Prav tako bi lahko prišlo do zapletov z izplačevanjem mikroplačil, saj so upravičeni stroški in način porabe javno financiranih projektov pri nas strogo regulirani.

Primeri dobre prakse – Tehnologija PyBossa je že bila uspešno uporabljena za množičenje v številnih raziskovalnih projektih na spletni platformi Crowdcrafting.org. Na spletni strani platforme[7] med uporabniki tehnologije PyBossa navajajo npr. Britanski muzej (British Museum), švicarski raziskovalni inštitut CERN in Združene narode (UNITAR).

Platformo smo že uspešno namestili in trenutno testiramo njeno funkcionalnost in možnosti oblikovanja množičenjskih kampanj. Prvi primeri mikronalog bodo na njem na voljo do konca leta 2015.

5 VLOGA MNOŽIČENJA PRI NAČRTOVANJU SLOVARJA SODOBNEGA SLOVENSKEGA JEZIKA

V nadaljevanju predstavljamo analizo potreb in preliminarne predloge, kako bi bilo mogoče množičenje uporabiti za obdelavo podatkov pri nekaterih fazah gradnje novega slovarja. Zasnova končnih mikronalog za množičenje bo močno odvisna od okoliščin slovarskega projekta, kot so višina in trajanje financiranja, partnerji in projektni načrt, zato v tem razdelku s primeri nalog le ponazarjamo nekatere možnosti uporabe množičenja v različnih fazah izdelave slovarja v scenarijih s krovnim financiranjem in nizkoproračunskih okvirih.

5.1 Leksikon besednih oblik

Analiza potreb in določitev ciljne skupine množičnikov: Leksikon besednih oblik ima znotraj projekta izdelave slovarske baze dve različni, a medsebojno prepleteni vlogi, ki pogojujeta tudi načrtovanje njene vsebine. V prvi vrsti je namenjen prikazovanju informacij o pregibnih, naglasnih, besedotvornih in drugih oblikoslovnih lastnostih slovarskih iztočnic. Druga, slovarskemu uporabniku skrita, a prav tako pomembna vloga leksikona besednih oblik pa je njegova

7 http://crowdcrafting.org/about (dostop 8. 8. 2015).

uporaba v različnih jezikovnotehnoloških aplikacijah za procesiranje slovenskega jezika, ki potrebujejo informacije o oblikoslovnih in izgovornih lastnostih slovenskega besedišča, kot so črkovalniki, pregibniki, razpoznavalniki in sintetizatorji govora, strojni prevajalniki itd.

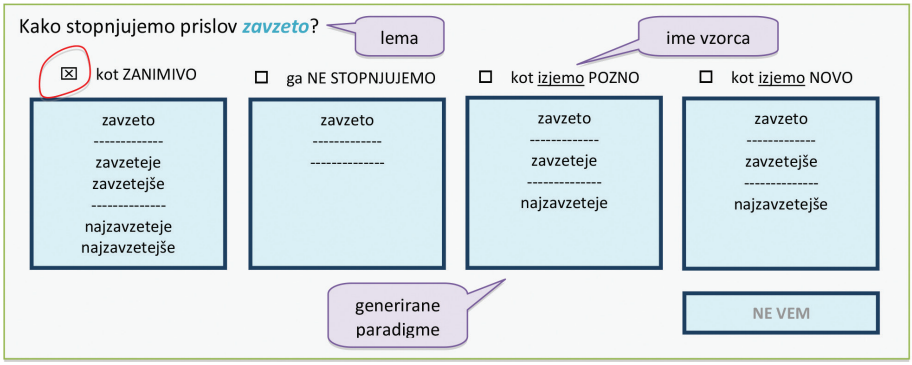
Za razliko od druge, jezikovnotehnološke vloge leksikona besednih oblik, ki v ospredje postavlja predvsem čim večjo pokritost besedišča, sta pri prvi, jezikovnopriročniški vlogi leksikona pomembna predvsem čim večja natančnost in zanesljivost prikazanih podatkov. To zahteva precejšnjo količino ročnega dela, zato bi prenos zamudnih rutinskih nalog z leksikografa na množičnike bistveno pospešil in izboljšal proces človeške validacije strojno izluščenih korpusnih podatkov. Čeprav so tovrstni problemi za avtomatske algoritme še vedno trd oreh, so za materne govorce slovenščine z jasnimi navodili in nekaj uvajanja razmeroma enostavni, zato bomo ta sklop nalog pripravili tako, da jih bo lahko reševal čim širši krog ljudi.

Oblikovanje mikronalog in zlatega standarda: Za potrebe priprave slovarske baze smo predvideli tri primere nalog: določanje standardne pregibne paradigme slovarske iztočnice, določanje standardnih besedotvornih povezav slovarskih iztočnic in razširitev leksikona besednih oblik za potrebe jezikovnih tehnologij. Eksperti bodo za vsako nalogo čiščenja leksikona izdelali ločen zlati standard. Za vsako besedno vrsto bo v zlatih standardih vključenih predvidoma po 200 primerov s treh frekvenčnih pasov iz korpusa: tretjina zelo pogostih (s frekvenco v korpusu Gigafida več kot 1000), tretjina srednje pogostih (s frekvenco med 1.000 in 100) in tretjina redkih (s frekvenco pod 100).

5.1.1 Določanje standardne pregibne paradigme slovarske iztočnice

Predvidene metode avtomatskega luščenja oblikoslovnih podatkov temeljijo na predpostavki, da sta že pred izdelavo leksikona na voljo ročno pregledana seznama slovarskih iztočnic s podatkom o besedni vrsti (leme geslovnika) in vseh standardnih vzorcev pregibanja v slovenskem jeziku. Za vsako iztočnico pregibnih besednih vrst so nato na podlagi seznama vzorcev in podatkov v referenčnem korpusu avtomatsko generirane možne oblikoslovne paradigme za dano lemo, množičniki pa med prikazanimi paradigmami[8] izberejo pravilno (Slika 3).

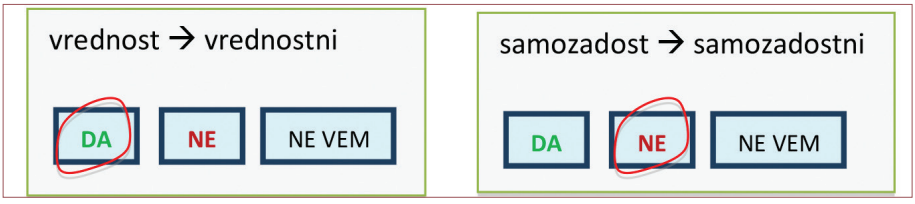
8 Pri tem med procesom avtomatskega luščenja (tj. iskanjem preseka generiranih paradigem in korpusnih podatkov) ter procesom množičenja dopuščamo možnost dodatnega avtomatskega filtriranja na obvladljivo število možnosti (npr. do največ tri paradigme), npr. s statističnimi izračuni verjetnosti glede na delež v rabi izkazanih oblik ali upoštevanjem informacij v leksikonu Sloleks, če je slovarska iztočnica vanj že vključena. Prav tako lahko prikazane paradigme razvrstimo od najverjetnejše do najmanj verjetne.



Slika 3: Mikronaloga za določanje pregibne paradigme, kjer uporabnik s klikom izbere eno izmed možnosti.

5.1.2 Določanje standardnih besedotvornih povezav slovarskih iztočnic

Leksikon poleg informacij o pregibanju slovarskih iztočnic prinaša tudi informacijo o njihovih besedotvorno povezanih oblikah znotraj vnaprej določenega nabora besedotvornih povezav (npr. med samostalnikom in izpeljanim svojilnim pridevnikom ali glagolom in deležnikom). Besedotvorne povezave slovarskih iztočnic z drugimi lemami, ki so obenem lahko (ne pa nujno) tudi same del slovarskega geslovnika, podobno kot pri luščenju pregibnih paradigem na podlagi vnaprej znanega nabora iztočnic in besedotvornih vzorcev iz korpusa luščimo avtomatsko. Za razliko od naloge v razdelku 5.1.1, kjer uporabniki izberejo eno izmed več prikazanih možnosti, za validacijo ustreznosti para izhodiščne in povezane leme, v tem pomenu predvidevamo klasično nalogo zaprtega tipa z odgovori da, ne in ne vem (Slika 4).

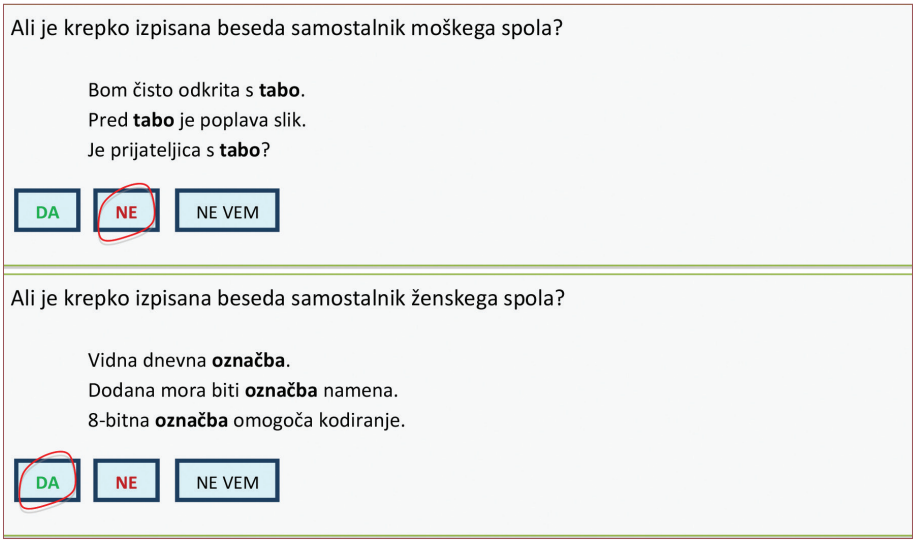


Slika 4: Primer uporabe množičenja za validacijo besedotvornih povezav.

5.1.3 Razširitev leksikona besednih oblik za potrebe jezikovnih tehnologij

Kot smo že omenili, jezikovnotehnološka vloga leksikona besednih oblik predvideva vključitev veliko širšega nabora lem in ne zgolj tistih, ki ustrezajo

iztočnicam v slovarju. Potencialne nove, v slovarski podmnožici leksikona nezabeležene leme je mogoče iz korpusa pridobiti z avtomatskimi metodami (npr. s strojnim lematizatorjem ali z besedotvornimi pretvorbami obstoječih lem), izluščene leme pa so nato skupaj s potencialnimi primeri rabe v osnovni obliki dane v presojo množici (Slika 5).



Slika 5: Primeri potencialnih samostalniških lem iz korpusa KRES, ki nista vključeni v izhodiščni leksikon Sloleks.

Za nadaljnje določanje pregibnega vzorca, besedotvornih povezav in morebitnih nestandardnih variant tako potrjenih novih lem lahko nato uporabimo enake metode množičenja kot pri izdelavi jezikovnopriročniške podmnožice leksikona besednih oblik, a jih je z metodološkega vidika smiselno osamosvojiti v ločeno, drugo fazo izdelave, saj lahko ročno validirano slovarsko podmnožico leksikona izkoristimo kot učno množico za izboljšanje postopkov avtomatskega luščenja korpusnih podatkov, stopnjo njihove ročne validacije pa lahko prilagajamo potrebam jezikovnih tehnologij, pri katerih je velikost pogosto pomembnejša od natančnosti.

Tretja zanimiva možnost izrabe množičenja za potrebe jezikovnih tehnologij pa je ročno razdvoumljanje besednih oblik v kontekstu na mestih, kjer strojni označevalnik zaradi več možnih interpretacij iste besedne oblike naleti na visoko stopnjo dvoumnosti.

Rekrutiranje množičnikov in upravljanje s kampanjo: Ker je prečiščen leksikon ključen za ostale faze izdelave slovarja, je pomembno, da je na njem v čim krajšem času opravljenega čim več dela, tako da si bomo v tem sklopu k množičenju prizadevali pritegniti čim širšo skupino množičnikov, ki nimajo nujno

jezikovne izobrazbe. Zato bomo pri teh kampanjah toliko več pozornosti posvetili temeljitemu uvajanju, urjenju in testiranju množičnikov s pomočjo vnaprej pripravljenih predstavitvenih videoposnetkov, vaj z avtomatizirano povratno informacijo o pravilnem odgovoru in strogem filtriranju nezanesljivih odgovorov in množičnikov. V teh kampanjah bomo veliko pozornosti namenili tudi ozaveščanju skupnosti o pomenu kvalitetnih, javno dostopnih jezikovnih virov, zato kampanjo načrtujemo tako, da bi množičnike motivirali z mesečnimi materialnimi nagradami za najuspešnejše sodelavce (vrednostnimi boni, vstopnicami).

5.2 Leksikalna baza

Analiza potreb in določitev ciljne skupine množičnikov: V okviru izdelave leksikalne baze so glavni izzivi, pri katerih bi si lahko pomagali z množičenjem, postaviti dobra izhodišča za pomensko členitev iztočnice v slovarju, narediti izbor relevantnih kolokacij in iz korpusa izluščiti učinkovite slovarske zglede. S pomočjo predlaganih nalog želimo v procesu množičenja ugotavljati pomenske povezave med korpusnimi zgledi, ki vsebujejo obravnavano besedo v določeni besednozvezni kombinaciji, in predlaganimi pomenskimi opisi zanjo.

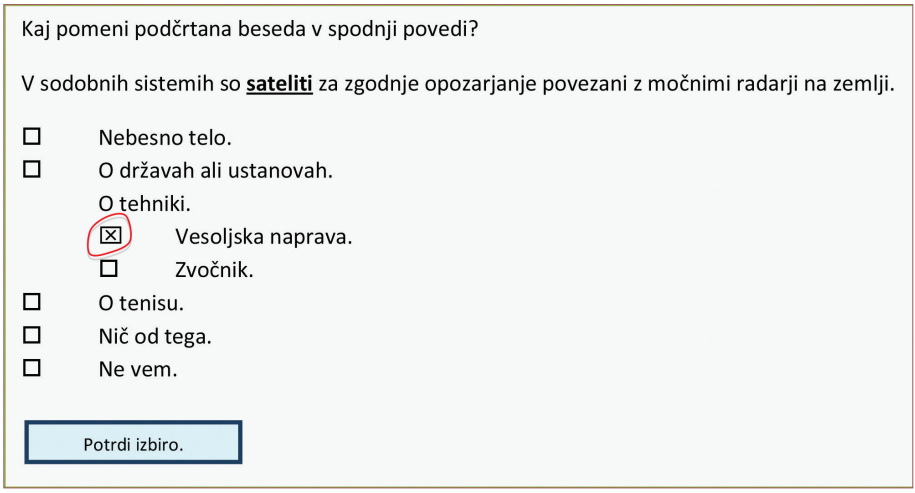
Dobra pomenska členitev je po našem mnenju taka, ki odraža čim večji konsenz v jezikovni skupnosti, zato predvidevamo, da bo prav analiza odgovorov, ki jih bodo prispevali množičniki nestrokovnjaki, omogočila prepoznavanje pomenskih opisov, ki so neproblematični in izkazujejo najvišjo stopnjo strinjanja, ter situacije, kjer je povezava med zgledom in pomenom razpršena. Nadaljnja analiza teh primerov s strani leksikografov bi omogočila izboljšavo pomenske informacije v slovarju tako z vidika pomenskega opisa kot z vidika stopnje pomenske razčlenjenosti.

Oblikovanje mikronalog in zlatega standarda: Podatki, iz katerih pri oblikovanju nalog za množičenje izhajamo v tem sklopu, so avtomatsko izluščeni iz korpusa Gigafida: prek orodja Sketch Engine so z uporabo funkcij Besedne skice in GDEX na podlagi skladenjskih struktur izluščeni kolokatorji za posamezne leme in zgledi, ki pripadajo kolokaciji (tj. zvezi kolokatorja + leme). Za zlati standard bodo uporabljeni podatki o pomenski členitvi in ustreznimi zgledi iz Leksikalne baze za slovenščino (LBS), ki so jo na podlagi korpusnih podatkov ročno izdelali izkušeni leksikografi.

5.2.1 Pripisovanje pomena

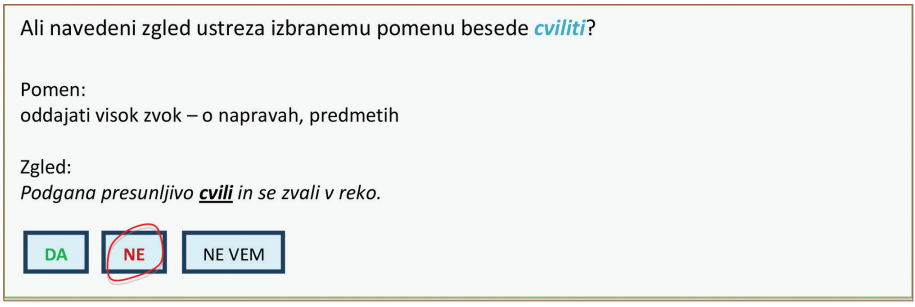
V prvi nalogi množičnikom ponudimo različne pomenske opise večpomenske besede, kot so zabeleženi v LBS, in jih prosimo, da zgled, v katerem je beseda

navedena v določeni kolokaciji, pripišejo pomenu, ki se jim zdi najbolj ustrezen. Pri tej nalogi lahko množičnik izbere samo en pomenski opis.



Slika 6: Mikronaloga za pripisovanje pomena besedi v kolokaciji.

Cilj druge naloge je enak, s tem da v tem primeru množičnikom ponudimo pomenski opis besede in jih prosimo, da ugotovijo, ali mu navedeni zgled, ki vključuje besedo v določeni kombinaciji, ustreza.



Slika 7: Mikronaloga za potrjevanje pripisanega pomena.

Rekrutiranje množičnikov in upravljanje s kampanjo: Ker pripisovanje pomena, razvrščanje kolokacij in identifikacija učinkovitih zgledov zahteva precej strokovnega znanja in izkušenj, si bomo v teh kampanjah k sodelovanju prizadevali pritegniti samostojne leksikografe, podiplomske študente in mlade diplomante jezikovnih smeri. Ker je količina dela, ki ga je v tem sklopu potrebno opraviti v čim krajšem času, velika, zahteve po natančnosti pa visoke, načrtujemo množičnike v teh kampanjah motivirati z mikroplačili. Da bomo rekrutirali zanesljive

množičnike, bomo poskrbeli s temeljitim predhodnim in sprotnim testiranjem sodelujočih.

5.3 Norma

Analiza potreb in določitev ciljne skupine množičnikov: V primeru, ko so osnovna ali pregibne oblike slovarske iztočnice povezane s pogosto jezikovno zadrego, želimo uporabnika ustrezno usmerjati z zanesljivimi in informativnimi podatki o njihovi normativni zaznamovanosti. Pri luščenju in obravnavi variantnih oblik bomo nadaljevali s konceptom, ki je bil razvit v okviru izdelave Slogovnega priročnika SSJ (Krek et al. 2013b) in v okviru katerega smo postopke množičenja že preizkusili kot pomoč pri pripisovanju normativnih podatkov pri tistih oblikah, kjer je pripis ustrezne normativne oznake in kategorije odvisen od izgovora, pomena ali drugih lastnosti besedišča, ki presegajo trenutne zmogljivosti strojnega procesiranja, za rojene govorce jezika pa predstavljajo razmeroma nezahtevno nalogo (K. Dobrovoljc in Krek 2013).

Ta sklop nameravamo v veliki meri vpeljati v študijsko prakso, saj se vsebinsko povezuje s predmetoma Uvod v študij slovenskega jezika in Leksika in slovnica slovenskega jezika v okviru študija Medjezikovnega posredovanja na Oddelku za prevajalstvo Filozofske fakultete Univerze v Ljubljani.

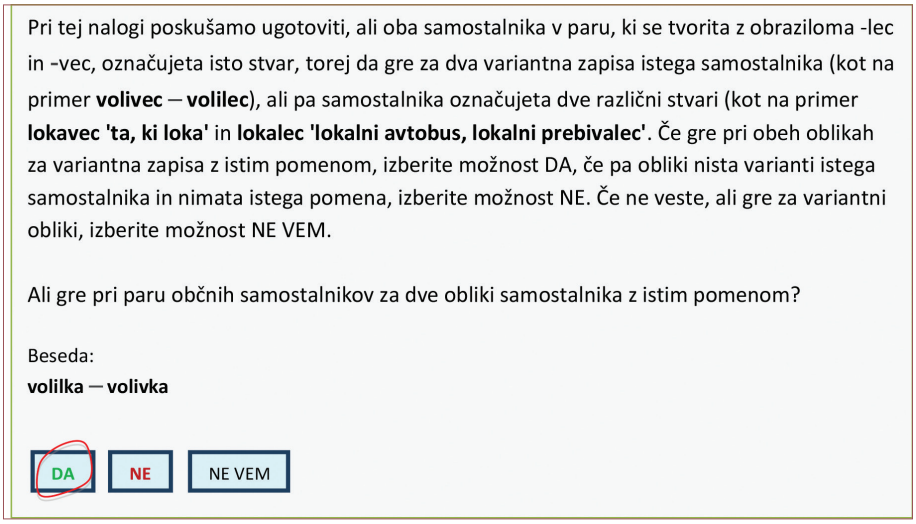
Oblikovanje mikronalog in zlatega standarda: V tem sklopu smo zaenkrat predvideli dve kampanji: določanje normativno zaznamovanih oblik in pregibanje tujejezičnih lastnih imen. Podatki za mikronaloge bodo na podlagi ročno določenih hevristik avtomatsko izluščeni iz korpusa, naloga množičnikov pa bo, da jih pregledajo in potrdijo oz. zavrnejo. Eksperti bodo za vsako od kampanj izdelali ločen zlati standard, ki bo vseboval po 300 reprezentativnih primerov.

5.3.1 Določanje normativno zaznamovanih oblik

Z izrazom normativno zaznamovane oblike označujemo vse tiste v rabi izkazane pregibne in besedotvorne besedne oblike, ki so normativno zaznamovane, nenevtralne oz. nestandardne (za nabor in argumentacijo kvalifikatorjev glej poglavja normativne in stilistične skupine). Eksplicitna kategorizacija variantnosti v pregibnih in besedotvornih paradigmah je pomembna, ker omogoča sistematično luščenje tovrstnih korpusnih podatkov (za razliko od dodajanja naključnih nestandardnih oblik), povezovanje z ustreznimi normativnimi pojasnili, prikazovanje opozoril o normativnih zadregah, povezanih z iztočnico,

na različnih mestih slovarskega sestavka in druge avtomatske analize, kot sta denimo priklic vseh normativnih zadreg ene iztočnice ali priklic vseh iztočnic z zadrego istega tipa.

Preden se lahko množičniki izrečejo o (ne)zaznamovanosti različic, jih morajo najprej prepoznati, kar ponazarja naloga na Sliki 8.

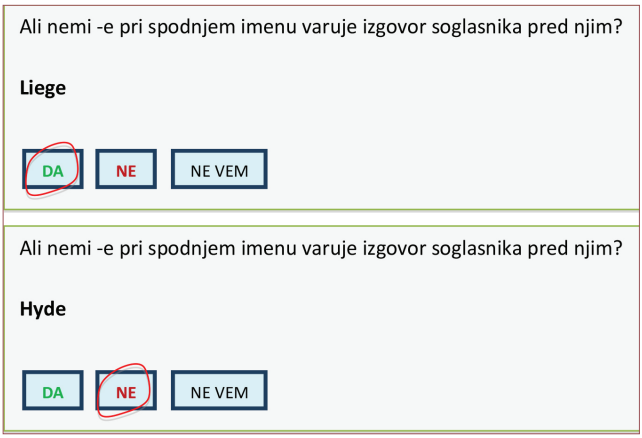


Slika 8: Iskanje parov variantnih lem znotraj kategorije D2c1a: Besedotvorje > Tvorba samostalnikov > Izbira priponskega obrazila > *-lec/-vec*.

5.3.2 Pregibanje tujejezičnih lastnih imen

Pregibanje tujejezičnih lastnih imen je drugi primer normativne jezikovne zadrege, pri kateri je uporaba množičenja tako rekoč nujna, saj je nemi -e na koncu tujeizvornih (lastnih) imen nepredvidljiv. Kadar nemi -e varuje izgovor soglasnika pred njim, ga pri pregibanju besede ohranjamo (npr. Wallace → Wallacea), kadar pa se izgovor soglasnika pred njim ob izpustu ne spremeni, tega pri pregibanju opuščamo (npr. Mike → Mika, Apple → Appla). Nemi -e na koncu besede lahko varuje soglasnike č, š, ž, dž in s (kadar je ta pisan s c) – primeri: Blanche → Blanchea, Limoge → Limogea, Dodge → Dodgea, Bruce → Brucea ipd.

Tokratna naloga množičnikov je, da s seznama avtomatsko izluščenih besed, ki se v korpusu Gigafida končujejo s t. i. nemim -e (npr. Gaye, Kaye ipd.), izberejo tiste, pri katerih se soglasnik pred njim izgovarja (Slika 9).



Slika 9: Primer mikronaloge za določanje imen, pri katerih nemi -e varuje izgovor soglasnika pred njim.

Rekrutiranje množičnikov in upravljanje s kampanjo: Glede na to, da nameravamo množičenjske kampanje vpeljati v študijsko prakso, bomo kvaliteto odgovorov zagotavljali z uvajalnimi predavanji in sprotnim preverjanjem njihovega razumevanja obravnavanega problema ter s pomočjo zlatega standarda. Načrtujemo, da bi študentske sodelavce motivirali z različnimi elementi družbene motivacije: poskrbeli bi, da med reševanjem poglobijo in nadgradijo svoje znanje slovenske slovnice in pravopisa, za sodelovanje bi jim priznali opravljanje obštudijskih obveznosti (prakse), prav tako pa bi jim izdali potrdilo o sodelovanju pri nacionalno pomembnem leksikografskem projektu, ki bi ga lahko kot referenco priložili k svojemu življenjepisu.

Čeprav množičenje ni najprimernejše orodje za normativnostna preverjanja v smislu anketiranja o preferenčnih jezikovnih sintagmah, si to možnost pridržujemo za morebitne jezikovne sklope, kjer bo zaradi slabih ali sploh neobstoječih jezikovnih podatkov normo potrebno vzpostavljati na tovrsten način.

5.4 Uporabniki

Analiza potreb in določitev ciljne skupine množičnikov: Čeprav se množičenje običajno ne uporablja za zbiranje subjektivnih ocen, je mogoče množičenjski sistem izkoristiti tudi na področju slovaropisnih uporabniških raziskav. Z določenimi prilagoditvami sistema je namreč mogoče vzpostaviti kontinuirano sodelovanje z ustrezno vzorčeno skupino (potencialnih) slovarskih uporabnikov, ki prispeva uporabniške evalvacije v zvezi s tehničnimi vidiki slovarja (iskalne možnosti, prikaz leksikalnih podatkov v vmesniku, značilnosti večpredstavnih vsebin ipd.).

Pri strukturiranju vzorca uporabnikov je treba upoštevati kategorije slovarske rabe in uporabniških skupin (Arhar Holdt 2015), kot tudi relevantne demografske značilnosti. Ocenjujemo, da bi za ustrezne posplošitve potrebovali v vzorcu vsaj 200 stalno sodelujočih. V poznejši fazi izdelave slovarja bi evalvacije lahko odprli tudi za splošno javnost in rezultate primerjali z odgovori fokusnih skupin.

Oblikovanje mikronalog in izdelava zlatega standarda: Uporabniško mnenje bo služilo kot podpora odločitvam na ravni slovarske vsebine, oblike in funkcionalnosti. Mikronaloge bodo v obliki izbire med dvema ali več možnostmi. Kot primer lahko podamo vprašanje s področja zapisa izgovora besede v slovarju. V vprašalniku nanizamo izvedbene možnosti (npr. različne vrste transkripcije ali različne možnosti dostopa do zvočnega posnetka), vprašani mora opredeliti, katera različica se mu zdi boljša (bolj uporabna, bolj intuitivna). Kot kažejo rezultati obsežnejših anketiranj slovarskih uporabnikov (Müller-Spitzer 2014), je pomembno ponuditi sodelujočim tudi možnost za odprte odgovore, kjer lahko podajo pojasnila glede svoje odločitve ali alternativne predloge.

Ker gre za nekonvencionalno uporabo množičenja, s katerim bomo preverjali mnenje in preference uporabnikov z metodami, ki so sorodne spletnemu anketiranju, za ta sklop nalog zlatega standarda ne potrebujemo.

Rekrutiranje množičnikov in upravljanje s kampanjo: Sodelujoči bodo rekrutirani iz splošne populacije, predvidoma s pomočjo inštitucij in društev, ki želeni uporabniški profil združujejo. Ob registraciji na množičenjsko platformo bo vsak sodelujoči izpolnil vprašalnik o izkušnjah, navadah in preferencah glede rabe slovarja. Sodelujoči bodo v sistem uvrščeni kot predstavniki določene uporabniške skupine (npr. lektor, prevajalec, učitelj slovenščine kot tujega/drugega jezika), s pomočjo uvodnega vprašalnika pa bo ta uvrstitev po potrebi dopolnjena. Z vprašalnikom bodo zbrane tudi dodatne informacije, pomembne za razumevanje vzorca.

Ko bo na voljo gradivo za evalvacijo, bodo sodelujoči na e-naslov prejeli vabilo k udeležbi z opredelitvijo, do kdaj je treba evalvacijo opraviti. Previdnost je potrebna predvsem pri številu evalvacij, ki ne smejo biti moteče in prepogoste, obenem pa ne preredke, da se ne izgubi vtis sodelovanja v skupini in motivacija za sodelovanje.

Kampanja mora biti zasnovana tako, da bo po koncu posamezne evalvacije prikazala statistike odgovorov z upoštevanjem zgoraj omenjenih kategorij. Na tak način bodo pripravljavci slovarja lahko hitro presodili, katera od predlaganih rešitev je najbolje sprejeta v celotnem vzorcu, kot tudi v posamezni uporabniški skupini. Statistični podatki morajo biti na voljo tudi pri vsakem posameznem članu, da je mogoč pregled nad odgovori skozi daljše časovno obdobje. S tem

vpogledom bi bil znan tudi delež neaktivnih, kar bi omogočilo pravočasno rekrutacijo novih sodelujočih.

6 ZAKLJUČEK IN PRIHODNJE DELO

Ob pravilnem načrtovanju in upoštevanju vseh ključnih načel oblikovanja in vodenja množičenjskih kampanj ni nobenega razloga, da množičenje pri gradnji slovarja sodobnega slovenskega jezika ne bi odigralo pomembne vloge pri zagotavljanju podpore leksikografom pri postprocesiranju šumnih avtomatsko izluščenih podatkov na ekonomsko in časovno vzdržen način ter z zanesljivimi rezultati. Kot je razvidno iz prispevka, smo za to pripravili vso potrebno organizacijsko, tehnično, vsebinsko in finančno podlago za učinkovito množičenje novega slovarja ter na podlagi tega predlagali optimalen delotok množičenja skupaj s ponazoritvami možnih množičenjskih kampanj v različnih fazah sodelovanja.

Izbrano platformo za množičenje smo že namestili, trenutno pa se pričenjajo priprave na temeljito testiranje predlagane metode, preizkušanje administratorskega in uporabniškega vmesnika, prilagajanje parametrov za zagotavljanje kvalitete ter nastavitev za uvoz in izvoz podatkov. V prihodnje nas čaka še identificiranje in razreševanje morebitnih dodatnih pravnih in logističnih preprek pri najemanju in plačevanju množičnikov ter seveda opravljanje pilotnih množičenjskih sej.

V specializiranih leksikografskih in jezikovnotehnoloških projektih je v minulem desetletju množičenje že postalo stalnica, spodbudni rezultati pa mu zadnja leta odpirajo vrata v vse večje in kompleksnejše leksikografske projekte nove generacije. Zato je pomembno, da njegov potencial izkoristimo tudi pri slovarju sodobnega slovenskega jezika in s tem postanemo referenčna točka za domače in tuje slovarske ter druge jezikovne vire.