PREVODOSLOVJE
IN UPORABNO
JEZIKOSLOVJE

Univerza v Ljubljani
FILOZOFSKA
FAKULTETA

Uredili: **Vojko Gorjanc, Polona Gantar, Iztok Kosem** in **Simon Krek**

SLOVAR SODOBNE SLOVENŠČINE: PROBLEMI IN REŠITVE

Uredili Vojko Gorjanc, Polona Gantar,
Iztok Kosem in Simon Krek

SLOVAR SODOBNE SLOVENŠČINE: PROBLEMI IN REŠITVE

Zbirka Prevodoslovje
in uporabno jezikoslovje

Ljubljana 2017

SLOVAR SODOBNE SLOVENŠČINE: PROBLEMI IN REŠITVE
ZBIRKA PREVODOSLOVJE IN UPORABNO JEZIKOSLOVJE
ISSN 2335-335X

Uredili: Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek
Recenzenti: Maja Bratanić, Wayles Browne in Václav Cvrček
Uredniški odbor zbirke: Špela Vintar, Vojko Gorjanc in Nike Kocijančič Pokorn
Tehnični urednik: Jure Preglau

Založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Izdal: Oddelek za prevajalstvo
Za založbo: Roman Kuhar, dekan Filozofske fakultete

Ljubljana, 2017
Prva izdaja, elektronska izdaja
Oblikovna zasnova: Kofein, d. o. o.
Prelom: Jure Preglau

Publikacija je brezplačna.
Publikacija je dostopna na: https://e-knjige.ff.uni-lj.si
DOI: 10.4312/9789612379759

Kataložni zapis o publikaciji (CIP) pripravili
v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID=292781824
ISBN 978-961-237-975-9 (pdf)

Kazalo vsebine

IV JEZIKOVNI VIRI ZA SLOVARSKI OPIS SODOBNE SLOVENŠČINE

V SODOBNA LEKSIKOGRAFIJA V TEORIJI IN PRAKSI

VI (IZ)GOVORJENO V SLOVARJU

VII SPECIALIZIRANA LEKSIKA IN SPLOŠNI SLOVAR

Predgovor

V monografiji *Slovar sodobne slovenščine: problemi in rešitve* so objavljeni rezultati raziskav, s katerimi želimo odgovoriti na nekatera ključna vprašanja pri snovanju koncepta sodobnega slovarja slovenskega jezika, ki bi bil kos izzivom sodobne leksikografije in bi (ponovno) umestil slovensko leksikografsko teorijo in prakso v mednarodni kontekst. Monografija prinaša 32 razprav, ki jih je prispevalo prav toliko avtorjev.

I

Pri pripravi koncepta novega slovarja sodobne slovenščine, ki bo temeljil na spoznanjih pričujoče monografije in hkrati nadgrajeval Predlog za izdelavo Slovarja sodobnega slovenskega jezika,[1] smo se raziskovalci pod okriljem Centra za jezikovne vire in tehnologije Univerze v Ljubljani[2] organizirali v Konzorcij za jezikovne vire in tehnologije. Pri tem je bila priprava izhodišč za slovarski koncept le eden od konzorcijevih namenov, saj je njegova vizija dolgoročno sodelovanje pri pri raziskovanju, vzpostavljanju in vzdrževanju jezikovnih virov in jezikovnotehnoloških orodij za sodobni slovenski jezik.

Pri vsebini dela smo izhajali iz dejstva, da mora biti namen novega slovarja dvojen: (1) informirati govorce slovenščine in vse druge uporabnike o leksikalno-slovničnih lastnostih slovenskega jezika na način, ki sledi sodobnim leksikografskim praksam; (2) zagotavljati temeljni leksikalni in slovnični vir za razvoj jezikovnotehnoloških aplikacij.[3] Da bi bil slovar čim bolj koristen tudi za jezikovnotehnološke namene, mora biti delo že v izhodišču zasnovano kot odprto dostopna računalniško procesljiva baza podatkov. Taka zasnova omogoča tudi izdelavo slovarskih opisov za različne ciljne skupine, saj se zavedamo potrebe slovenskega prostora po pripravi jezikovnih virov, ki bodo zadovoljevali različne uporabnike, od šolajoče se mladine do jezikovnih profesionalcev, kot so npr. prevajalci in uredniki, od govorcev slovenščine kot maternega oziroma prvega jezika do tistih, ki se slovenščine učijo kot tujega jezika. Zato je bil eden od namenov raziskav tudi odgovoriti na vprašanje, kakšna so uporabniška pričakovanja glede slovarskih informacij in kako jim zadostiti.

Celotno delo smo zasnovali v okviru tematskih skupin, raziskave pa usmerili ciljno, glede na odprta vprašanja, ki jih zastavlja slovarski projekt za novo dobo in spremenjene okoliščine (globalnega) komuniciranja. Ker mednarodni trendi kažejo, da je digitalni, predvsem spletni medij, tako klasični kot mobilni, tisti, v katerega je treba vlagati vso energijo pri pripravi jezikovnih opisov, je to tudi

1 http://www.sssj.si/ (dostop 7. 8. 2015).
2 http://www.cjvt.si/ (dostop 7. 8. 2015).
3 Več o tem na http://www.cjvt.si/projekti/sssj/ (dostop 7. 8. 2015).

izhodišče, na katerem gradimo svoje delo. Izhajamo iz take zasnove slovarja, kjer so slovarske informacije na voljo uporabnikom že v procesu nastajanja, hkrati pa se ves čas vsebinsko in pojavno prilagaja medijem, ki se vse bolj uporabljajo za prenos jezikovnih informacij uporabnikom. Zato je ključnega pomena, da jezikoslovje pri raziskavah in razvoju jezikovnih opisov v digitalnih medijih tesno sodeluje s strokovnjaki z različnih področij, saj se le tako lahko skupaj načrtuje in razvija sodobnim medijem vsebinsko in pojavno prilagojene jezikovne opise. Pričujoča monografija je rezultat prav tovrstnega interdisciplinarno zasnovanega raziskovalnega dela.

Delo na monografiji je potekalo v glavni koordinaciji Vojka Gorjanca in s koordinatorji tematskih področij: Špele Arhar Holdt (slovarski uporabniki), Kaje Dobrovoljc (leksikon), Darje Fišer (množičenje), Polone Gantar (leksikografija), Roberta Grošlja (slovnica), Petra Jurgca (fonetika in izgovarjava), Monike Kalin Golob (stilistika in kvalifikacija besedišča), Simona Kreka in Marka Robnik-Šikonje (računalniška podpora: zaledni del, predstavitveni del in luščenje), Nataše Logar in Tomaža Erjavca (korpusni viri in zapis korpusa), Damjana Popiča (norma), Marka Stabeja (sociolingvistična umestitev), Darinke Verdonik (govorjeni jezik), Špele Vintar (terminologija) in Gregorja Žaklja (večpredstavnost). Skupaj je pri delu tematskih skupin sodelovalo 49 strokovnjakov z različnih področij: jezikoslovja, računalništva in oblikovanja.

Koordinatorji tematskih skupin so delo organizirali na različne načine in z različno dinamiko, pri delu vsake od tematskih skupin pa smo sodelovali člani glavnega projektnega odbora: Polona Gantar, Vojko Gorjanc, Iztok Kosem, Simon Krek in Marko Robnik-Šikonja. Na ta način je bilo delo, ki se je v tematskih skupinah logično tesno medsebojno prepletalo, usklajeno. Po vmesnem delovnem srečanju koordinatorjev februarja letos, na katerem je bila predstavljena vsebina dela v posamezni skupini, so koordinatorji na koncu poskrbeli tudi za uskladitev razprav s svojega področja za objavo v tej monografiji.

II

Monografija prinaša razprave v desetih tematskih sklopih: (1) Slovar umeščamo v slovenski sociolingvistični kontekst načrtovanja korpusa slovenščine, ob tem pa programsko izpostavimo ključne točke spremenjene ideologije, ko gre za vprašanje standardnojezikovne kulture. (2) Predstavljamo delo na sodobnem slovarju, ki ni zgolj tehnološko podprto, ampak v izhodišču jezikovnotehnološko zasnovano, z vključevanjem sodobnih postopkov analize jezikovnih podatkov, njihovega hranjenja in prikaza. Ob tem pa razprave naslavljajo tudi vprašanja sodobnih standardizacijskih praks in prinašajo predloge rešitev za slovenščino. (3) V slovenskem

prostoru se prvič podrobneje lotevamo analize slovarskih uporabnikov in njihovih potreb, po vzoru uporabniških raziskav v evropskem prostoru. Programsko se odmikamo od nedefiniranih in notranje nestrukturiranih konceptov, kot so splošni, običajni, povprečni, zahtevni uporabnik, in jih nadomeščamo z analizo slovarske rabe in jasno definiranih slovarskih uporabnikov. (4) Ker se korpusni viri za sodobne jezikovne opise logično vedno znova premišljajo, saj tako razvoj jezikovnih tehnologij kot tudi nenehno spreminjanje diskurznih praks v določenem jezikovno-kulturnem prostoru narekuje tudi spremenjen odnos do jezikovnega vira, predstavljamo tovrstne razmisleke v zvezi z osrednjim korpusnim virom, tj. nadgradnjo korpusa Gigafida. (5) Kompleksen leksikografski postopek od korpusne analize, tudi z uporabo besednih skic za slovenščino, avtomatskega luščenja jezikovnih podatkov in pridobivanja dobrih slovarskih zgledov do uporabe moči množic v določenih fazah leksikografskega procesa itd. predstavimo kot izhodišče leksikografskih razmislekov. Ti se med drugim navezujejo na sodobno tujo, zlasti evropsko leksikografsko teorijo in prakso, pri čemer poleg anglosaškega posebej predstavljamo tudi francoski prostor. (6) Razpravljamo o tem, kako v slovarskem opisu pristopiti do govorjenega jezika in se posvečamo naglasu in izgovarjavi. Pri slednjem se načrtno odmikamo od jezikoslovno usmerjenega modela k opisu, ki ima v ospredju slovarskega uporabnika. (7) V današnji družbi znanja ima v diskurzih pomembno vlogo specializirana leksika, zato tudi temu segmentu slovarja posvečamo posebno pozornost, in sicer tako z uporabniškega vidika kot tudi jezikovnotehnološkega, v okviru katerega predstavljamo luščenje specializirane leksike za splošni slovar. (8) Predstavljamo predlog stilističnega označevanja v slovarju, pri čemer se odmikamo od dosedanjega zvrstnega modela in ga nadomeščamo z analizo diskurznih praks. Zanima nas, v katerih primerih se določena raba uveljavlja in je na ravni jezikovne skupnosti prepoznana kot stilno, časovno, vrednotenjsko ali kako drugače specifična ter kako to informacijo vključiti v različne segmente geselske zgradbe. (9) V sklopu slovničnih razprav je ključnega pomena nov predlog besednih vrst v slovenskem jeziku, saj dosedanji slovarski opisi izkazujejo neusklajenost, zato je za sklop jezikovnih opisov, ki jih načrtujemo, pomembna jasna in konsistentna besednovrsna opredelitev, ki vzdrži slovarsko aplikacijo. (10) Monografijo zaključuje sklop razprav o uporabi moči množic. Predstavljamo tako pregled področja množičenja in njegov potencial v sodobni leksikografiji kot tudi konkreten predlog, ki določa način in namen uporabe tega postopka v različnih fazah izdelave slovarja sodobnega slovenskega jezika.

III

Ker je slovar – kot smo na začetku že omenili – logično načrtovan za digitalni medij, nekateri segmenti dosedanjega dela, npr. celostna grafična podoba slovarskih projektov, večpredstavnost ipd., v tej monografiji niso objavljeni, preprosto

zaradi narave medija, tj. tiskane knjige. Na voljo bodo v elektronski knjigi, ki bo izšla spomladi leta 2016. Tam bodo nekatere rešitve, ki v tiskanem mediju ne morejo biti dovolj prepričljivo predstavljene, prikazane tudi na novemu mediju prilagojen način.

Raziskave in rešitve, ki jih prinaša monografija, so zasnovane širše od samega koncepta slovarja, seveda pa bodo, kot je bilo načrtovano, vključene v končni koncept Slovarja sodobnega slovenskega jezika. Brez raziskovalno široko zasnovanega leksikografskega dela do vrste rešitev preprosto ne bi prišli, zato je pomembno, da se dinamika leksikografskega raziskovalnega dela v slovenskem prostoru, ki je do sedaj izkazoval velik zaostanek za sodobno leksikografijo v svetu, tudi v prihodnje nadaljuje.

Zasnova dela na tej monografiji v okviru Konzorcija za jezikovne vire in tehnologije je pokazala, kako lahko uspešno zastavimo ciljno usmerjene raziskave, hkrati pa so raziskovalci izkazali željo in potrebo po timskem delu in odprti vsebinski diskusiji tudi v prihodnje. Načrt urednikov je, da skupaj s sodelavci vzpostavljeno sodelovanje ohranimo, ga nadgradimo in v tem okviru usmerjamo nadaljnje slovenistično raziskovalno delo.

Uredniki
Ljubljana, Porto in Herstmonceux,
avgust 2015

I
Sodobni slovar slovenskega jezika

Daj mi slovar in spremenim ti (jezikovno) skupnost

Marko Stabej

Abstract

This paper places lexicography in both the historic and contemporary contexts of status and corpus planning for Slovenian. The critical conclusion reached is that status planning is a very generalised concept, through which it is not possible to clearly define planned and actual processes in a language community, whereas corpus planning and language resources are closely related to directing social relations within a community. Since the end of the 18th century, lexicography in Slovenia has had an increased symbolic value; this remains the case today, especially in monolingual lexicography. Analysis of the role of the state as the main driver behind modern language policy reveals that the state, for the most part, does not take responsibility for facilitating the planning of language resources for Slovenian, leaving this instead to the field of Slovenian Studies, and expecting a united stance on the issue. However, there is a wide range of disparate views on language and the nature and purpose of lexicography in this field, and relations in the language community are very different – to the point where there is an ideological divide. Our ideological credo in general is that the discourses of all Slovenian speakers of the Slovenian language are equal and worthy of note.

Keywords: language planning, language policy, lexicography, state, linguistic ideology

Ključne besede: jezikovno načrtovanje, jezikovna politika, slovaropisje, država, jezikoslovna ideologija

Jezik je cena, ki jo je treba plačati za realnost,
in ta narod je ne bo mogel kupiti,
dokler bo javna raba slovenščine planska.

Marko Crnkovič, 1992

1 UVOD

236 let mineva, odkar si l. 1779 je v prvem zvezku almanaha Pisanice alegorična Krajnska dužela zaželela svoj dikcijonarjum. »/D/a tega nimam jest, tu je še mene sram,«[1] pravi zaskrbljeno v enem od dvanajstercev.[2] Potrebovala ga je za svojo čast in slavo in zato vljudno, a odločno prosila svoje sinove, naj ji ga naredijo in pri tem pozabijo na lastno korist in medsebojne razprtije. Dežela Kranjska je seveda v tekstu alegorija, avtorjeva projekcija – in Legiša (1977: 425) brez dlake na jeziku pravi, da je bila ta pesem »nekaka reklamna napoved dela, ki ga je pripravljal podjetni sobrat Pohlin in je izšlo l. 1781 kot Tu malu besediše«. Tako vzvišena prošnja reklama? Da – saj če smo pravični, je prava dežela Kranjska kakšen slovar s slovenščino l. 1779 že imela, tudi v tiskani, kaj šele rokopisni obliki, vendar se vsaj tej alegorični deželi iz pesnitve ti prejšnji slovarji niso zdeli dovolj dobri za potešitev njenega ponosa in dokaz vrednosti njenega jezika. Zato je potrebovala novega, pravega, ki bi dokazal, da je njen jezik samostojen, samonikel in bogat in da tudi po jeziku ne zaostaja za sosedami in kolegicami ... Če pa se neusmiljeno odpovemo iluziji, gre ravno za obratno razmerje: novi, skoraj že pripravljeni Pohlinov trojezični slovar je potreboval prepričljiv in močan raison d'être, potreboval je družbeni ugled, potreboval je, da ga nekdo pomemben nujno potrebuje ... in kdo (poleg morda matere Cerkve, ki pa pri posvetni razsvetljenski dejavnosti, tudi če so jo izvajali duhovniki, ni bila več na prvem mestu) je bil tedaj lahko družbeno bolj pomemben kot političnoupravna dežela sama?

Najbrž ste opazili, da nekdo pri zgornjem zapletu manjka, po čemer lahko sklepamo, da ni prav zelo pomemben člen: to so govorci in govorke oz. jezikovna skupnost. Slovar je potrebovala dežela za svoj ponos, ne ljudje (ali celo ljudstvo) za svoje potrebe. Iz tega in iz drugih zgodovinskih zgodb lahko izpeljemo misel, da je med tem, kdo slovar potrebuje (ali se komu zdi, da kdo ga), in med tem, kdo ga (naj) uporablja, bistvena razlika (z mnogo vmesnimi odtenki). Še danes se rado zgodi, da govorcev in govork pri razpravljanju o jezikovnih priročnikih in virih ni oziroma so kvečjemu prikrito vsebovani globoko v kaki metonimiji. Marko Snoj (Perdih 2009: 5) tako npr. pravi: »Slovenska strokovna in laična javnost namreč

1 Besedilo navajamo po Legiševi izdaji Pisanic (1977).
2 Analiza in kontekstualizacija besedila Krajnska dužela želi tudi svoj dikcijonarjum imeti v Stabej (2009a: 126–129).

soglašata, **da naš jezik nujno potrebuje nov slovar** (poudaril M. S.), saj je od izida zadnje knjige Slovarja slovenskega knjižnega jezika minilo že 17 let, od pripravljalnih in konceptualnih del zanj pa že pol stoletja, poleg tega se današnji čas razlikuje od časa nastajanja tega slovarja po družbenem in ekonomskem redu, kar je deloma vplivalo tudi na vrednostni sistem naroda, in po bliskovitem vzponu sodobnih tehnologij.«

Kdo torej potrebuje kakšen slovar, kako ga narediti, kdo, kako, zakaj in s kakšnim namenom naj ga uporablja? Vse to so nenehno odprta vprašanja – ki si jih je treba pri raziskovanju in snovanju slovaropisne dejavnosti redno postavljati, hkrati pa nanja po najboljših močeh in vesti vsaj začasno odgovoriti, če si želimo kak slovar pripeljati do končnega izdelka (kar v sodobnih informacijskih, tehnoloških in družbenih okoliščinah pomeni bistveno manjšo stopnjo statičnosti kot v nekdanjih tiskanih časih).

2 SLOVAR V JEZIKOVNONAČRTOVALNEM OKVIRU

Umestimo sodobno slovaropisno dejavnost v okvir jezikovnega načrtovanja in njegove dialektike. Kot je znano, se načrtujejo status, korpus in znanje jezika oz. jezikov (Jernudd in Nekvapil 2012: 24). Načrtovanje korpusa jezika pomeni v običajni predstavi dejanje standardizacije in nato vzdrževanje standardiziranosti nekega jezika. Izdatno načrtovanje korpusa postane zares potrebno šele takrat, ko naj bi s tem dosegli take ali drugačne družbene učinke; prva stopnja je ponavadi homogenizacija jezikovne skupnosti, ta pa omogoča nadaljnje učinke, npr. povečanje dostopnosti izobrazbe, povečanje stopnje različnih oblik družbene participacije, lahko pa tudi prevlado neke verske ali etnične oz. narodne skupnosti, s tem tudi ideologije in posledično prerazporeditev družbenih elit in podobno – ti učinki oz. cilji se lahko odkrito ali prikrito združujejo in prepletajo. Načrtovanje korpusa posameznega jezika je zato običajno pojmovano kot nujni pogoj za načrtovanje oziroma doseganje ustreznega statusa tega jezika v govorno večjezičnih družbah – in ponavadi gre za jezik skupine ljudi z izhodiščno pretežno nizkim socialno-ekonomskim statusom. Toda – kaj pa je ustrezen status jezika? In nenazadnje, kaj status sploh je in po čem se meri? Status je položaj nekega jezika v skupnosti in v družbi, bi bila najosnovnejša parafraza – ki pa prikriva vsaj toliko, kot odkriva. Stereotipno o rasti slovenskega jezika in njegovega statusa razmišljamo kot o premočrtnem procesu vsaj od prve slovenske knjige l. 1550 do ustave samostojne Slovenije l. 1991. Kot da s(m)o si vsi Slovenci (in Slovenke) vsaj od prve polovice 16. stoletja do l. 1991 nenehno in močno želeli živeti samo v slovenščini; ko da so nam to dolga leta preprečevali samo sovražni tujci, verski, posvetni oblastniki in drugi krivci, ki so nam našo naravno pravico

grobo odrekali. Pri natančnejšem pregledu slovenske sociolingvistične zgodovine hitro spoznamo drugačno, manj zvezno in vzneseno sliko zgodovinskih predstav o slovenščini in njeni vlogi. Slovenski protestantje si kljub svoji vročični jezikovni dejavnosti nikoli niso predstavljali slovenščine tudi kot edinega jezika tedanje posvetne oblasti ali jezika izobraževanja ali jezika znanosti, saj bi bilo to v takratnih evropskih razmerah nekaj nepredstavljivo nenavadnega ali kar čudaškega; če ne bi bilo tako, ne bi predvidevali npr. v svojih stanovskih šolah kazni za tiste, ki si (razen v prvem razredu) drznejo spregovoriti v slovenščini (ali kakšnem drugem maternem jeziku), saj je za izobrazbo štela samo latinščina (Ahačič 2007: 18). Prvič lahko željo po slovenskem življenju samo v slovenščini najdemo v političnem programu Zedinjena Slovenija l. 1848, pa še to je pravzaprav interpretacija; Matija Majar, eden od avtorjev programa, je v posebnem letaku želeni status slovenščine v okviru narodnega programa opredelil takole: a)[3] »slovenski jezik mora imeti v slovenskih krajih popolnoma to pravico, ktero ima nemški v nemških, italijanski v italijanskih …« b) »mora nam biti svobodno (frei) upeljati slovenščino v vse pisarnice in šole višje in nižje v Slovenii … ako hočemo in kakor hočemo« c) »vsaki uradnik (Beamter) v slovenskih krajih, kteri se za naprej bode postavil v službo mora popolnama slovensko znati.« (Granda 1999: 96) Kako lahko tako načrtovani status ustrezno opredelimo z eno samo besedo ali besedno zvezo? Tako sociolingvistika kot pravna nomotehnika seveda imata rešitve, od skrajno zbanaliziranega hierarhiziranja na »m-jezike« in »v-jezike« (Toporišič 1991: 143), do poimenovanj uradni ali državni jezik in podobno (Vidovič Muha 2003; Stabej 2015). Toda tako v zgodovini kot v sedanjosti družbenega trenutka je dejanski (in načrtovani) status jezikov v družbi in jezikovni skupnosti (pa tudi pri posameznikih in posameznicah) nekaj zelo kompleksnega.[4] Ves ta ekskurz je pravzaprav namenjen podkrepitvi naslednje teze: če drži, da je načrtovanje korpusa nujno povezano z načrtovanjem statusa, moramo torej vedeti, kakšen status jezika načrtujemo, da znamo zasnovati in izpeljati ustrezno načrtovanje korpusa. Načrtovalci najbrž ne vedo zmeraj natanko in v vseh podrobnostih, za kakšen status jezika si prizadevajo, in tudi med načrtovalci in v skupnosti sami se predstave o pričakovani statusni razporeditvi jezikov in drugih zaželenih elementih jezikovne situacije lahko opazno razlikujejo, ne da bi se tega sami jasno zavedali. Lahko si npr. vsi želijo, da bi njihov jezik postal uradni ali državni – vendar jim to pomeni zelo različne stvari; različna razmerja med pravicami in dolžnostmi zasebnih govorcev, javnih in državnih ustanov v javni komunikaciji, v različnem obsegu in različni dinamiki uresničevanja pravic ipd. Seveda mora načrtovanje korpusa – ali širše pojmovano načrtovanje jezikovne

3 Oznake pripisal M. S.

4 Kompleksnost načrtovanja statusa ponazorimo še z enim zgodovinskim zgledom: Matija Majar Ziljski (kot tudi mnogo drugih izobražencev istega časa, pa tudi pred in po njem) je ob vsem jasnem zavzemanju za utrjevanje in dvig statusa slovenščine razvijal tudi precej drugačen hkratni scenarij, ki ga le post festum lahko označimo za neke vrste »plan B«: postopno združevanje južnoslovanskih jezikov v en standardni jezik; zdelo se je namreč verjetneje, da bo le tak jezik s svojim množičnim zaledjem lahko statusno dovolj močan povezovalec slovanstva v konkurenci drugih močnih jezikov z velikim številom govorcev (Stabej 2010: 136).

opremljenosti nekega jezika in jezikovne skupnosti – izkazovati pravo mero robustnosti v razmerju do heterogenosti predstav o načrtovanem statusu. Sledi lahko v grobem le povprečenemu trendu načrtovanega statusa, saj bi bila sicer zlasti v manj številnih in manj premožnih jezikovnih skupnostih jezikovnonačrtovalna dejavnost v obliki slovarjev, slovnic in drugih priročnikov, prilagojena diverzificiranim pogledom na nadaljnjo podobo jezikovne skupnosti (in torej na načrtovani status), takorekoč nemogoča. Toda – če poenostavimo – tudi vso slovensko zgodovino so bile odločitve o tem, kakšne slovnice, slovarje in druge podobne reči naj se izdelujejo, povezane s takimi ali drugačnimi sporazumevalnimi, kulturnimi in drugimi potrebami, oziroma natančneje, s takimi ali drugačnimi projekcijami, kaj naj bi bilo za koga ali kaj je najbolj potrebno – ter seveda s stvarnimi izvedbenimi možnostmi. Prva slovaropisna dejanja v zvezi s slovenščino so bila pretežno praktično usmerjena – Register v Dalmatinovi Bibliji 1584 je imel za nalogo premoščati regionalne razlike v besediščni kompetenci bralcev pri razumevanju svetopisemskih besedil; Alasijev Vocabolario Italiano, e Schiauo (1607) je bil namenjen začetnemu jezikovnemu pouku slovenščine ter posledično lažjemu delu bodočih italijansko govorečih menihov devinskega samostana med (samo) slovensko govorečim ljudstvom; Megiserjevi slovarski deli, ki sta vsebovali slovenščino, Dictionarium quatuor linguarum (1592) in Thesaurus Polyglottus (1603) sta že merili širše, zadnje je v predgovoru avtor namenjal: »ne samo filologom in študentom jezikoslovja /.../ ampak tudi profesorjem katerihkoli drugih umetnosti in znanosti, /.../ zlasti pa še zgodovinarjem, geografom, zdravnikom, kemikom, knežjim pisarjem in odposlancem in vsem drugim, ki mnogo potujejo, trgovcem in preprodajalcem dišav, pa tudi vojakom in vojaškim poveljnikom« (J. Stabej 1977, prevod K. Gantar). Nimamo ne prostora ne potrebe za nadaljnji zgodovinski pregled povezav med slovaropisno dejavnostjo in načrtovanjem statusa slovenščine v slovenskem prostoru; omenimo le, da gre od omenjenega Pohlinovega slovarja l. 1781 naprej za vse bolj zapletena razmerja, saj se v dotlej pretežno praktično slovaropisno delovanje vse močneje vpletajo tudi simbolne razsežnosti. Nenazadnje o tem priča saga o nastajanju velikega slovensko-nemškega slovarja v 19. stoletju, ki je bila po neštevilnih zapletih z nejasnostjo zasnove in ciljev, negotovostjo financiranja in posledično hudimi izvedbenimi težavami zaključena šele s Pleteršnikovim Slovensko-nemškim slovarjem (1893–1895). Najmočnejšo simbolno vlogo v slovenski jezikovni skupnosti (in njej primerljivih) pa ima tisti slovar, ki se zgodovinsko-razvojno gledano pojavi najkasneje: enojezični.

3 ENOJEZIČNI SLOVAR, ENOJEZIČNI VIR

Tradicionalni splošni enojezični slovar večjega obsega je sam po sebi vir vseh drugih priročniških in drugih jezikovnoinformativnih izpeljav s svojo domnevno

referenčnostjo (saj temelji na prepričanju, da slovar opisuje, kakšen jezik res je).[5] Proces nastajanja slovarja je sicer lahko bolj ali manj natančno opisan, ohranjeno ali celo digitalizirano je lahko tudi listkovno gradivo, ki je bilo podlaga leksikografskega dela (ponavadi pa je le zelo omejeno dostopno). Vse to pa ne spremeni dejstva, da enojezični slovar v skupnosti obvelja kot merodajni vir informacij o jeziku.[6] Referenčnost je njegova temeljna vloga; prav zato v zgodovinskem procesu oblikovanja slovenske (ali katerekoli druge »zamišljene« enojezične skupnosti) od dvo- ali večjezičnih slovarjev prevzame tudi najvišjo simbolno vlogo v jezikovni skupnosti. Enojezični slovar je namreč simbolni dokaz, da je neka jezikovna skupnost zrela za samostojnost – za enojezičnost pač. Hkrati enojezični slovar to skupnost opredeljuje in usmerja, načeloma z dvema glavnima vzvodoma: a) kaj izbere za vredno uslovarjenja b) na kakšen način in s kakšnim namenom izbrano uslovari. Usmerjanje lahko poteka v obliki sugeriranja (ali celo predpisovanja) »pravih«, družbeno zaželenih jezikovnih praks in odsvetovanja (ali celo prepovedovanja) »nepravih«, družbeno nezaželenih; ker so jezikovne prakse zmeraj družbenega značaja, s tem slovar oblikuje tudi družbena razmerja. Seveda lahko enojezični slovarji glede omenjenih glavnih vzvodov uberejo zelo različne poti, kot nenazadnje z najrazličnejših vidikov podrobno predstavlja tale knjiga. Čisto temeljno vprašanje pa vendarle ostaja naslednje: se sodobna slovaropisna dejavnost z odločitvijo za enojezični slovar kot svoj glavni cilj vendarle izhodiščno ne omeji usodno in preveč? So glavni argumenti za tako odločitev vsebinske ali simbolne narave? Najbrž po nepotrebnem zaostrujemo vprašanje: argumenti so obojne narave. Simbolno je smiselno delati enojezični slovar zato, ker ima v družbi večjo težo in ker se zanj glede na pretekle izkušnje veliko verjetneje lahko pridobi javni denar. To je sicer banalen in pragmatičen, a še kako tehten razlog. Po drugi strani je veliko tudi vsebinskih razlogov. Običajna sodobna uporabnojezikoslovna predstava je, da je dober in kakovosten enojezični slovarski vir najboljša podlaga za izdelavo dvo- in večjezičnih virov (čeprav je, kot smo omenili, zgodovinsko gledano pravzaprav obratno) in ima torej izdelava enojezičnega slovarja pri gradnji jezikovne opremljenosti prednost pred dvojezičnimi. Malo manj očiten vsebinski argument za izdelovanje enojezičnega slovarja pa je prav možnost rekonceptualizacije jezikovne skupnosti, jezika samega, jezikovnega znanja, pa tudi jezikoslovne dejavnosti in še marsičesa s tem povezanega – kar je vse veliko bolj sistematično uresničljivo (ali sploh možno) pri snovanju enojezičnega slovarja oz. vira, kot bi bilo pri dvo- ali večjezičnem.

5 Domnevnost pomeni avtorsko skonstruiranost po eni strani in privzemanje te vrednosti pri jezikovni skupnosti po drugi strani, ki tega ne počne iz dejanskega lastnega dojemanja jezikovne referenčnosti in reprezentativnosti samega dela in njegovih konkretnih, posameznih slovarskih rešitev, ampak pretežno iz družbene avtoritete ustanove avtorjev/izdajateljev/založnikov takega referenčnega dela.

6 Z razvojem korpusnega jezikoslovja in prosto dostopnostjo referenčnih korpusov se odpirajo možnosti za spreminjanje take prakse, saj lahko slehernik podatke o avtentični jezikovni rabi pridobi sam; toda to možnost izkoristiti ni nekaj trivialnega, nasprotno, brez temeljitih premikov v izobraževanju, pa tudi v jezikovnih in siceršnjih stališčih ostaja možnost predvsem na teoretični ravni, takorekoč neizkoriščena.

Referenčnost temeljnega sodobnega enojezičnega (hipo)slovarja (če lahko s hiposlovarjem poimenujemo idealno urejeno, torej z vsemi potrebnimi podatki opremljeno in po vseh poteh dostopno slovarsko bazo) npr. ni več nujno (samo) v tem, da iz množice vseh diskurzov v nekem, npr. slovenskem jeziku in/ali iz introspekcije jezikovne zmožnosti avtorjev rigorozno izbere in uslovari tisto, kar naj bi bilo v tem jeziku najboljše, najbolj vredno in kar naj bi ta jezik posledično edinole bil, temveč da zajame in evidentira vse diskurze, ki jih lahko označimo recimo za slovenske. S tem jezikovni vir ne izenačuje dejanske (četudi skonstruirane) družbene vrednosti različnih pojavnih oblik jezika in diskurza, in če jih opisno označuje, jih še ne razvršča po njihovi vnaprejšnji splošni vrednosti, ampak – po možnosti na čimbolj nevtralen način – opredeljuje njihovo vlogo v družbi oz. jezikovni skupnosti, razen seveda v tem, da jih vse razglaša za del slovenščine. To ima lahko zelo verjetno za čisto praktično posledico večjo in drugačno povezanost slovenske jezikovne skupnosti, ki je ne bi povezovala več le reprezentativna oblika knjižnega jezika, temveč tudi vse druge jezikovne prakse, od – če speljemo misel do skrajnosti, kar je praktično kljub vsem komaj verjetnim teoretskim, metodološkim in tehnološkim izboljšavam pri jezikoslovnem delu, ki jih lahko podrobno spoznavamo v tejle monografiji, še vedno neuresničljivo – najbolj odročnih krajevnih govorov do najbolj hibridnih oblik elektronsko posredovanega sporazumevanja, od začetniškega jecljanja tujcev, ki se šele uvodoma spoznavajo s slovenščino, do fosilizirane slovenske govorice starejših generacij slovenskih izseljencev v Argentini. Če bi bili na nekem mestu vsi ti diskurzi zajeti skupaj in vsi njihovi jezikovni izrazi opredeljeni za sestavine slovenščine, bi bil to precejšen korak naprej od medlega zavedanja o slovenskem jezikovnem zvrstnem kontinuumu, ki pretežno živi samo med slovenistično stroko in znanostjo, pa še to zares živo in brez vrednotenjske stereotipiziranosti le med nekaterimi. Pri mejnem odločanju o slovenskosti (pravzaprav bi morali govoriti o »slovenščinskosti«, o lastnosti, da je nekaj v slovenščini) diskurzov v jezikovnem smislu torej na tej ravni, recimo ji najsplošnejša, namesto nekdanjega izrazito izključevalnega nabora meril lahko uporabimo obratno skrajnost – vsevključevalnost. Ne bi torej iskali in določali, kaj je še slovensko, ampak kaj je že slovensko. Tak obrat lahko seveda beremo kot golo retoriko, ki ne prinaša nobene vsebinske razlike. Vendar ne gre le za retoriko, ampak za stvarno namero. Pri določanju kriterijev bi se bilo seveda treba opreti ne le na korpusne podatke o različnih diskurzih, temveč tudi na empirično raziskovanje procesov razumevanja in delovanja jezikovne zmožnosti. Če se za trenutek vrnemo k razpravi o razmerju med načrtovanjem statusa in korpusa: sodobni enojezični slovar mora najbrž biti zasnovan v okviru redefinicije sodobne slovenske jezikovne situacije. Ta je po eni strani zaznamovana s statusom slovenščine kot uradnega jezika Republike Slovenije in uradnega jezika Evropske unije, po drugi strani pa z diverzifikacijo jezikovnih praks v slovenščini in z izrazito povečano večjezičnostjo v vseh ozirih (Stabej 2015). Kakšno naj bo načrtovanje korpusa v tem kontekstu in kaj želimo s tem doseči?

4 DRŽAVA KOT JEZIKOVNA NAČRTOVALKA JEZIKOVNE OPREMLJENOSTI

Odnos slovenske države kot najmočnejše jezikovnopolitične instance do slovarske problematike (in tudi siceršnje problematike jezikovnih virov in jezikovne opremljenosti) kaže tole grobo sliko:

Od osamosvojitve naprej država bolj ali manj stalno in stabilno (=neprojektno) financira ustanovo, ki naj bi za to domnevno sistematično skrbela, torej Inštitut za slovenski jezik Frana Ramovša Znanstvenoraziskovalnega centra Slovenske akademije znanosti in umetnosti (odslej ISJFR ZRC SAZU), a brez kakršnegakoli pravega in jasnega pričakovanja o ciljih ter rezultatih, kaj šele nadzora nad tem – o tem natančno poroča na podlagi omejenih javno dostopnih podatkov Simon Krek (2014).

Prvi slovenski nacionalni program za jezikovno politiko 2007–2011[7] je sicer vključeval tudi slovarsko opremljenost slovenščine, toda precej jasno je, da ta ni sodila med njegove prednostne jezikovnopolitične naloge. Slovarje skupaj z »učbeniki, jezikovnimi korpusi ipd.« program prišteva v temeljno jezikovno infrastrukturo in »delno opremljenost s temeljno jezikovno infrastrukturo« celo navaja med »ugodnostmi« pri »ovrednotenju razmer za jezikovno strategijo«, čeprav obsega in narave te infrastrukture nikjer ne opredeli. Kot vzporedno »slabost« tedanjih razmer navaja, da »izpopolnjevanje temeljne jezikovne infrastrukture ne poteka sistematično in usklajeno (jezikovni viri pogosto niso širše dostopni raziskovalnim skupinam in temeljno jezikovno razvojnim podjetjem).« To je gotovo (bil) resničen problem, toda hkrati izkazuje prepričanje, da je jezikovna infrastruktura namenjena predvsem specialistom (znanstvenikom, raziskovalcem in poslovnim razvojnikom) – ne pa neposredno celotni jezikovni skupnosti. Slovarje resolucija omenja le nekajkrat, izraziteje na treh mestih: kot podcilj č) 11. cilja (za višjo sporazumevalno kulturo) navaja »zagotavljanje spletne dostopnosti jezikovnih virov, npr. SSKJ, SP in drugih«; podcilj l) je »priprava splošnih in specializiranih priročnikov za slovenščino«, kamor po programu sodijo »frazeološki, sinonimni, terminološki, zgodovinski in dvojezični slovarji«; nazadnje še v podcilju m) »izpopolnjevanje in zagotavljanje spletne dostopnosti elektronskih jezikovnih orodij«, kar naj bi po programu pomenilo tele naloge: »črkovalnik, prevajalniki, slovarji, terminološke zbirke«. Da slovarske naloge v tedanjem programu niso imele skoraj nobene teže, lahko sklepamo tudi po zanje predvidenem denarju; za podcilja č) in m) je bil ocenjen strošek 83.458 evrov na vsakega, za podcilj l) pa 104.323 evrov. Skupaj torej 271.239 evrov, od česar naj bi pri l) in m) po 41.729 evrov prišlo iz neopredeljenih »dodatnih zasebnih virov«, torej naj bi iz proračuna Republike

7 Resolucija o nacionalnem programu za jezikovno politiko 2007–2011 (ReNPJP0711). Uradni list RS, št. 43/2007, 18. 5. 2007, str. 5952 (https://www.uradni-list.si/1/content?id=80272, dostop 22. 7. 2015).

Slovenije za širše slovarske zadeve v okviru jezikovnopolitičnega programa v petletnem obdobju namenili 187.781 evrov. Glede na to, da je skupno program za pet let predvideval porabo 12.705.616 evrov proračunskega denarja, je torej slovarski dejavnosti namenil vsega 1,48 odstotkov vseh sredstev.

Na podlagi navedenega lahko posredno izpeljemo nekaj sklepov:

1. Program izkazuje prepričanje, da je bila tedanja temeljna slovarska opremljenost slovenščine – s Slovarjem slovenskega knjižnega jezika in Slovenskim pravopisom – načeloma čisto zadovoljiva, potrebno je bilo poskrbeti le za njeno večjo (spletno) dostopnost; nekaj malega naj bi bilo treba nadgraditi le posebne, specializirane slovarske informacije, zato »frazeološki, sinonimni, terminološki, zgodovinski« in čisto na koncu sramežljivo tudi »dvojezični slovarji«, čeprav je že iz finančne konstrukcije, pa tudi zaradi drugih indicev popolnoma jasno, da program vsaj s tem zadnjim ni mislil čisto resno.
2. Tako malo namenjenega denarja tem dejavnostim lahko pripišemo tudi prepričanju avtorjev resolucije, da se potrebno redno obnavljanje slovarske opremljenosti za slovenščino itak plačuje iz drugih virov, in sicer že prej omenjeni domnevno samoumevno za te zadeve pooblaščeni ustanovi, ISJFR ZRC SAZU. Ta je – posredno sodeč po programu – ostala še naprej edina pristojna tudi za določanje dinamike temeljnega slovarskega opremljanja slovenskega jezika, in sicer tistega, ki simbolno največ šteje: enojezičnega. To posredno potrjuje tudi dvodnevni Strokovni posvet o novem slovarju slovenskega jezika l. 2008 (Perdih 2009), ki sta ga priredili SAZU in ZRC SAZU v zaprtem krogu povabljenih strokovnjakov in strokovnjakinj, pobudo zanj pa je dal in ga gmotno omogočil prav Sektor za slovenski jezik Ministrstva za kulturo Republike Slovenije; ministrstvo je tudi plačalo izid zbornika s posveta.
3. Vse to posredno pomeni, da po resoluciji realnega načrtovanja in izvajanja slovarske opremljenosti slovenščine v Republiki Sloveniji ne more početi nihče drug kot omenjena ustanova – saj ni mogoče pričakovati, da bi to dejavnost lahko plačeval kdor koli drug kot država.[8]

A država ni le eno ministrstvo, tudi če je to edino, ki je v izvršni veji oblasti RS neposredno pristojno za slovenski jezik in jezikovno politiko v zvezi z njim; de facto ima bistveno močnejše jezikovnopolitične vzvode od Ministrstva za kulturo (vsaj na področju načrtovanja jezikovne zmožnosti in oblikovanja jezikovnih

8 Zadnji obsežnejši neproračunsko financirani slovarski izdelek na Slovenskem je bil DZS-jev Veliki angleško-slovenski slovar OXFORD, ki je izšel v dveh zvezkih 2005–2006 (Grabnar in Šorli 2008). Zanimivo je, da je ta slovar v svojem zelo kratkem pozdravnem nagovoru na omenjenem strokovnem posvetu omenil akademik Janez Orešnik: »Na malo bolj lokalni ravni pa bi omenil nekaj, kar se najbrž že pozablja, in to je, da smo v zadnjem času dobili tudi zajetno slovarsko delo, ki ni nastalo na Inštitutu za slovenski jezik Frana Ramovša, je pa seveda slovenistično, in to je slovenski del Velikega angleško-slovenskega slovarja« (Perdih 2009: 11) – morda kot svarilo pred nadaljnjo izgubo prevladujočega slovaropisnega položaja lastne ustanove?

stališč) ministrstvo, pristojno za izobraževanje oz. šolski sistem. Dinamika reorganiziranja ministrstev slovenske vlade ravno na teh področjih je bila v zadnjih letih precejšnja. Do 2012 so bila to tri ministrstva: za kulturo; za šolstvo in šport; za visoko šolstvo, znanost in tehnologijo. 2012 je 10. slovenska vlada vsa tri ministrstva združila v eno Ministrstvo za izobraževanje, znanost, kulturo in šport, 11. vlada pa je 2013 spet osamosvojila ministrstvo za kulturo in ohranila ministrstvo za izobraževanje, znanost in šport, česar tudi aktualna vlada z nastopom l. 2014 ni spremenila.[9] Od 2008 do 2013 je potekal projekt Sporazumevanje v slovenskem jeziku,[10] ki sta ga financirala Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za izobraževanje, znanost in šport, ki je tudi formalno skrbelo za obsežen projekt petih konzorcijskih partnerjev. Projekt, oziroma v evrojeziku »operacija« je potekala v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013, z razvojno prioriteto »razvoj človeških virov in vseživljenjskega učenja«, prednostna usmeritev pa je bila »izboljšanje kakovosti in učinkovitosti sistemov izobraževanja in usposabljanja 2007–2013«. Projekt omenjamo zato, ker je priskrbel vrsto ključnih elementov sodobne jezikovne opremljenosti slovenščine,[11] ki se danes zdijo, vsaj nekateri – kot se vidi tudi iz naše knjige – popolnoma nepogrešljivi tako v teoretičnem kot uporabnem slovenističnem jezikoslovju, še posebej pri slovaropisni dejavnosti (od spektra najpotrebnejših korpusnih virov in jezikovnotehnoloških aplikacij do poskusnih verzij uporabni(ški)h jezikovnih in jezikovnodidaktičnih priročnikov za slovenščino). Takorekoč nič od tega ni nastalo v okviru nacionalnega jezikovnopolitičnega programa ne s spodbudo vladnega organa, neposredno pristojnega za slovenščino, ne na ustanovi, ki naj bi bila strokovno najbolj pristojna za tovrstno načrtovanje, torej ISJFR ZRC SAZU, ampak po drugih poteh ter z drugimi akterji. Podatke o viru in načinu financiranja izdelave sodobne jezikovne opremljenosti slovenščine v tem obdobju lahko pretvorimo tudi v zgodbo s pomenljivo prispodobo: jezikovna opremljenost ni bila izdelana kot izrecno jezikovni cilj v okviru uradnega nacionalnega jezikovnopolitičnega programa, temveč kot eden od vzvodov nadaljnjega povezovanja in izboljševanja skupnosti (zato jo je sofinanciral Evropski socialni sklad) ter razvoja človeških virov.[12]

Novi jezikovnopolitični program Republike Slovenije je nastajal od l. 2011 in bil po razmeroma zapletenem ciklusu 15. 7. 2013 sprejet v Državnem zboru v obliki

9 Podatki o reorganizacijah vlade so povzeti po Wikipediji: (https://sl.wikipedia.org/wiki/9._vlada_Republike_Slovenije; https://sl.wikipedia.org/wiki/10._vlada_Republike_Slovenije; https://sl.wikipedia.org/wiki/11._vlada_Republike_Slovenije; https://sl.wikipedia.org/wiki/12._vlada_Republike_Slovenije (dostop 26. 7. 2015).

10 http://www.slovenscina.eu/projekt (dostop 26. 7. 2015).

11 Videopredstavitev rezultatov z zaključne konference projekta na http://videolectures.net/zakljucnakonferencassj2013_ljubljana/ (dostop 23. 7. 2015). Problemom infrastrukture slovenščine in slovenistike je bil posvečen zbornik simpozija Obdobja 28 (Stabej 2008), tudi spletno dostopen.

12 Zadnja besedna zveza kar kliče po sarkastični družbenokritični interpretaciji (človeški viri = nekaj, kar omogoča kapitalu in kapitalskim elitam nenehno bogatenje), pa se bomo temu odrekli; razvoj človeških virov razumimo idealistično kot povečevanje življenjske moči in družbene vključenosti posameznikov in posameznic.

Resolucije o nacionalnem programu za jezikovno politiko 2014-2018.[13] Na njeni podlagi sta v organizaciji Ministrstva za kulturo nastala tudi usklajena osnutka dveh t. i. akcijskih načrtov, za jezikovno izobraževanje in za jezikovno opremljenost.[14] Ne moremo v podrobnosti, izražamo pa mnenje, da so tako veljavna resolucija kot oba osnutka akcijskih načrtov strokovno bistveno modernejši, konsistentnejši, uporabnejši in tudi bolje medinstitucionalno usklajeni jezikovnopolitični dokumenti od prejšnje resolucije.[15] V njih je našla svoje mesto tudi slovarska opremljenost slovenske jezikovne skupnosti, funkcionalno razčlenjena in trdno umeščena v širši kontekst druge potrebne jezikovne opremljenosti. Toda oba akcijska načrta sta – strokovno usklajena po javni debati – že od jeseni 2014 v ministrskem predalu. To skupaj z nekaterimi drugimi (ne)dejanji državnih organov kaže na indiferentnost izvršne oblasti v Republiki Sloveniji v zvezi z nadaljnjim razvojem jezikovnih virov in jezikovne opremljenosti, na kar je 26. 3. 2015 opozorila tudi izjava *Slovenščina in jezikovna antipolitika* l. 2014 ustanovljenega Konzorcija za jezikovne vire in tehnologije.[16] Državna letargija glede omogočanja (naročanja in financiranja) tovrstne dejavnosti vsaj zaenkrat nikakor ne pomeni, da na tem področju ni živahne dejavnosti, prav nasprotno. Znanstvena, strokovno-organizacijska, pa tudi lobistična in marketinška dejavnost (nekaterim dogajanjem bi lahko pripisali celo značaj obveščevalne in protiobveščevalne dejavnosti) vsaj od objave Predloga za izdelavo slovarja sodobnega slovenskega jezika, ki so ga napisali »trije jezikoslovci na podlagi dela na Leksikalni bazi za slovenščino v okviru projekta Sporazumevanje v slovenskem jeziku« (Krek et al. 2013), je skoraj vročična. Nimamo namena natanko opisovati dogajanja (čeprav bo to najbrž nekoč nujno – in zahtevno ter mučno opravilo), dodamo naj le to, da je konec marca 2015 ISJFR ZRC SAZU objavil svoj Osnutek koncepta novega slovarja slovenskega knjižnega jezika[17] in ga dal v enomesečno javno razpravo. V zagovor države in njenega jezikovnopolitičnega angažmaja na tem področju lahko rečemo, da je po objavi omenjenega prvega osnutka slovarja v letih 2013 in 2014, tudi na podlagi besedila jezikovnopolitičnega programa (še kot osnutka in kot že veljavnega dokumenta) ministrstvo za kulturo poskušalo doseči vsaj neko stopnjo strokovnega konsenza glede oblike in vsebine nadaljnje temeljne slovaropisne dejavnosti za slovenščino (npr. s prireditvijo posveta o novem slovarju slovenskega jezika 12. 2. 2014,[18] z vrsto formalnih in polformalnih pogovorov med

13 O temeljnih mejnikih nastajanja resolucije na spletni strani Službe za slovenski jezik pri Ministrstvu za kulturo: http://www.mk.gov.si/si/delovna_podrocja/sluzba_za_slovenski_jezik/resolucija_o_nacionalnem_programu_za_jezikovno_politiko_20142018/ (dostop 23. 7. 2015).

14 Akcijski načrt za jezikovno opremljenost je dostopen na: http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/Akcijska_nacrta/Akcijski_nacrt_za_jezikovno_opremljenost_javna_razprava.pdf (dostop 23. 7. 2015).

15 Pri tem gre seveda za subjektivno mnenje, saj sem bil avtor tega prispevka tudi vodja delovne skupine za pripravo osnutka jezikovnopolitičnega programa l. 2011 in vodja delovne skupine za pripravo osnutka akcijskega načrta za jezikovno izobraževanje l. 2013/14. Toda oznako se da vendarle podpreti tudi z objektivnejšimi kazalci, od tega, koliko ljudi, s katerih ustanov in na kakšen način je pri nastajanju sodelovalo, do tega, kako so vsebine programov strukturirane in povezane ipd.

16 Javno dostopna v obliki peticije na http://www.pravapeticija.com/jezikovna_antipolitika (dostop 23. 7. 2015).

17 http://www.fran.si/novi-sskj (dostop 27. 7. 2015).

18 Prispevki v obliki e-zbornika na http://www.mk.gov.si/si/delovna_podrocja/sluzba_za_slovenski_jezik/predstavitev_podrocja/dogodki/posvet_o_novem_slovarju_slovenskega_jezika/ (dostop 27. 7. 2015).

glavnimi strokovnimi akterji in predstavniki ter predstavnicami vpletenih ustanov), toda z novo vlado l. 2014 je politična volja glede tega očitno popolnoma usahnila. Da v »stroki« tudi ob mediaciji države ni prišlo do poenotenja pogledov na to, kakšna naj bo nadaljnja slovaropisna dejavnost v zvezi s slovenščino, ni pravzaprav nič čudnega – in pričakovanje, da se bo to prej ali slej zgodilo, je precej naivno. Slovaropisna nesoglasja so seveda prepletena tudi s čisto osebnimi in institucionalnimi razlogi vseh vpletenih in ti se še krepijo, dlje ko nesoglasja trajajo. Toda konsenz ali vsaj kompromis ni mogoč ne zaradi razlike v osebnostih, ampak zaradi (skoraj) nepremostljive razlike v konceptih – zato je bila zgoraj ob zadnji omembi stroka tudi v narekovajih. Zdi se namreč, da so pogledi na jezik, jezikovno skupnost, jezikovno situacijo, jezikovno načrtovanje in na globlji kontekst vsega tega v slovenskem in slovenističnem jezikoslovju zelo različni, še več – ideološko polarizirani. To ne pomeni politične polarizacije (tudi če si jezikoslovne ideologije kdaj skušajo utreti pot v financirano realnost s pomočjo strankarske politike ali si jih stranke občasno prisvojijo za svoje volilne ali druge politične potrebe), ampak razlikovanje v pojmovanju (jezikovnega) sveta. In kaj naj bi bile glavne razlikovalne ideološke poteze? Poskusimo nemogoče, ubesedimo tisti pol, s katerim se bolj ali manj identificiramo v tejle knjigi: pojmovanje izhodiščne enakovrednosti vseh govorcev in govork v jezikovni skupnosti (zato je raziskovanja vredna vsa jezikovna praksa in jezikovna zmožnost vseh govorcev in govork, pa tudi vsi diskurzi); posledično spoštovanje jezikovne različnosti v vseh oblikah; spoštovanje volje, potreb, želja in okusov govorcev in govork v njihovih jezikovnih praksah in jezikovnih izbirah; operacionalizacija skupnega knjižnega oz. standardnega jezika in izrecno raziskovanje njegovih ideoloških razsežnosti – in še bi morali naštevati ... ideološke razlike je namreč težko predstaviti izrecno, ker v izrecni obliki skoraj ne obstajajo; obstajajo le razlike v raziskovalnih metodologijah in praksah, v raziskovalnih ciljih ipd. Če se bo hotela država odgovorno odločiti o nadaljnjem razvoju opremljenosti slovenskega jezika in slovenske jezikovne skupnosti, je dobro, da te ideološke razlike razume in se jih pri odločitvah zaveda. Odločanje pa je odgovornost države, ne stroke (saj oba pola stroke ravnata ravno iz prepričanja o svoji odgovornosti). Prepričanje, da mora država pred odločitvijo počakati, da se stroka glede teh vprašanj poenoti, je pravzaprav le neodgovorno sprenevedanje, je le izogibanje odgovornosti. Ne odgovornosti do jezika: odgovornosti do skupnosti.

5 IDEOLOŠKI VIDIKI JEZIKOSLOVNE IN SLOVAROPISNE DEJAVNOSTI

Sodobno na korpusnih (in drugih relevantnih in referenčnih) jezikovnih podatkih temelječe in vsaj načelno družbeno neizključevalno slovensko jezikoslovje hoče biti čimbolj »opisno« – torej hoče izluščiti »slovenski jezik« iz statistično reprezentativnega zajema kar najširšega nabora različnih diskurzov različnih govork

in govorcev, ki so (konsenzualno) v slovenščini; osvoboditi se hoče (vsaj slovenske) jezikoslovne strukturalistične manire (po njej so podatki o jezikovni rabi le potrebna vžigalna svečica, ki da jezikoslovcu ali jezikoslovki iskrico za zagon teoretske mašinerije, ta pa nato, z obilno pomočjo nezanesljivega samoopazovanja njenega strojnika, konstruira t. i. jezikovni sistem, ki se razrase v avtonomno in brezprizivno instanco, v Sistem – in na koncu koncev je prava jezikovna raba lahko le tista, ki je v skladu s Sistemom, kar pa ni v skladu z njim in se med govorci vendarle rabi, je razglašeno le za anomalijo, in to ne le jezikovno, ampak tudi družbeno). Ta »nova« jezikoslovna paradigma (ki v slovenističnem jezikoslovju ni brez bogatega predhodništva) se skuša izogibati prezgodnjim poenostavljenim posplošitvam glede jezikovne regularnosti, se dosledno naslanjati na relevantne podatke, zmanjšati in dokumentirati postopke dekontekstualizacije jezika ter biti pri tem družbeno občutljiva in odgovorna. Teoretsko gledano je sicer to jezikoslovje izrazito hibridno, morda včasih celo teoretsko površinsko – pa si pravzaprav tudi ne želi apriorne teoretske izčiščenosti ali trdne navezanosti na to ali ono jezikoslovno »šolo«, navezuje se veliko bolj na svoje konkretno poslanstvo v konkretnih jezikoslovnih nalogah. Snovanje slovarja je gotovo tak katalizator konvergence in sinteze jezikoslovnih prizadevanj; sodobno jezikoslovje je namreč sicer (kar ni v okviru sodobnega delovanja znanstvenih disciplin in strok nič posebnega) izrazito specializirano in parcializirano. Konvergentnost slovaropisne dejavnosti pa pomeni tudi odpiranje prostora za interdisciplinarnost. Vse to se ne zgodi kar samodejno – za to si je treba dejavno prizadevati, to pa lahko počnemo le, če svojo dejavnost temeljito reflektiramo in se izogibamo vnaprejšnji samoumevnosti in domnevni danosti (kar še ne pomeni popolnega zanikanja tradicije).

Sama izbira opisnosti kot temeljne raziskovalne perspektive in uporaba gradivne podlage pri raziskovanju še ne pomenita odločilne razlike med jezikoslovnimi pristopi. Razlika se vzpostavi pri odločitvah, kaj, zakaj in kako naj opisujemo. Kako ideološko občutljivi so jezikoslovni postopki, tudi kadar se sklicujejo na empiričnost, si oglejmo ob naslednjem odstavku:

> Dokler ne bo slovensko jezikoslovje normiralo slovenske knjižne govorice na podlagi gradiva, se pravi posnetkov kultiviranih javnih govorcev, ljudi, ki po eni strani obvladajo hkrati z vsebino tudi knjižnojezikovno normo kot tako rekoč podzavestno sredstvo svoje besede, po drugi strani pa tudi psihološko breme javnega nastopa, do takrat bo normiranje slovenske javne (knjižne) govorice neprepričljivo (Vidovič Muha 1992).

Gre za perspektivo, ki je izključevalna in hkrati zelo interpretativno odprta, parafrazirajmo in problematizirajmo: relevantno gradivo za prepričljivo (za koga in po kakšnih merilih?) normiranje javne slovenščine (zakaj je pridevnik knjižne v oklepaju?) mora jezikoslovje najti pri govorcih, ki govorijo javno, sproščeno (če s tem prislovom dovolj pokrijemo njihovo »obvladanje psihološkega bremena

javnega nastopa«), ki so kultivirani (po starem SSKJ: »ki ima/jo/ splošno veljavnim načelom, normam, pravilom ustrezajoče lastnosti«), ki hkrati obvladajo vsebino (katero, česa; vsega, o čemer govorijo, svojega področja?) in knjižnojezikovno normo (kot podzavestno sredstvo svoje besede; ali to pomeni, da pri knjižnem govorjenju zvenijo naravno?). Interpretativna odprtost izjave ustvarja prostor za jezikoslovčevo oz. jezikoslovkino avtoritarno dejavnost (pri kateri mu oz. ji morda lahko pomagata tudi kak psiholog ali sociolog, toda glavne odločitve opravi on oz. ona). Odločiti pa se mora vsaj o tem: kdo je kultiviran, kdo obvlada vsebino in jezikovno normo, kdo obvlada javni nastop. Morda lahko jezikoslovcu priznamo izvedensko prednost pri odločanju o tem, ali je neka jezikovna praksa v skladu z veljavno knjižnojezikovno normo ali ne. Toda o kultiviranosti, prepričljivosti in všečnosti javnih govorcev imamo pravico soditi čisto vsi, ki jih poslušamo (in gledamo) in gotovo so sodobni okusi glede tega zelo različni, tudi v soodvisnosti od vrste diskurza, načina medijske posredovanosti ter drugih kontekstualnih dejavnikov – kar je bilo morda res v začetku 90-ih let v slovenskem medijskem prostoru še manj izrazito in prikrito, z razmahom t. i. komercialnih medijev pa je to zelo težko spregledati (Verovnik 2013). Obravnavani odstavek ustvarja vtis, kot da so kriteriji za uvrstitev govorcev med tiste, ki naj bodo vir norme, nekaj jasnega, neproblematičnega in enovitega. Navedek pa je tudi paradoksalen s stališča jezikovnega načrtovanja, vsaj na prvi pogled: nadaljnje normiranje jezika naj torej poteka na jezikovnih produkciji tistih govorcev, ki normo jezika že dobro obvladajo. Toda če normo dobro dejavno obvladajo, so se jo morali nekje naučiti. Naučili so se jo najbrž deloma izrecno med šolanjem, deloma pa so jo usvajali ob svoji receptivni in produktivni jezikovni dejavnosti med otroštvom in odraščanjem. Dobra norma torej že obstaja, saj brez nje ti govorci ne bi bili to, kar so – ali pa so ti govorci s svojo posebno dodano vrednostjo prej pusto in neživljenjsko jezikovno normo šele prekvasili v nekaj užitnega in uporabnega? Drugi paradoks je družbene narave: jezikovna produkcija posebno dobrih govorcev naj bo torej vir za govorno normo, ki bo prepričljiva. Ali prepričljiva pomeni taka, ki jo bodo sčasoma za svojo dejavno javno jezikovno rabo privzeli tudi vsi drugi govorci? Razen za izrazito radikalne pristaše popolne družbene enakosti se zdi ta misel absurdna; povprečiti hočeš torej posebej kakovostne govorce (ki ponavadi so ali vsaj naj bi bili tudi nosilci družbenega prestiža), povpreček ponuditi kot standard celi jezikovni skupnosti in s tem skupnost postopno jezikovno oz. govorno izenačiti? Kot da je razlika v jezikovnih praksah samo nekaj začasnega in prehodnega, ne pa nekaj, kar je nujni sestavni del zapletenih družbeno-psiholoških procesov človeške skupnosti. Kot da si tisti z nadpovprečnim družbenim prestižem želijo postati povprečni – tudi po jeziku.[19] Taka jezikovna enakost je enakost s figo v žepu, je kot jezikovna farma, kjer so vsi enaki, le nekateri so bolj enaki od drugih. Taka

19 Če že privzamemo, da si podpovprečni zmeraj želijo napredovati v povprečne; a spomnimo se Twainovega Huckleberryja Finna, kako je kot začasni posvojenec vdove Douglas trpel v dostojni obleki in čistem posteljnem perilu in rednih obrokih in nasploh urejenem ter pospravljenem življenju, v primerjavi s svojo siceršnjo svobodo.

jezikovna norma je prej neke vrste priročnik za neenakost – in je dober način jezikovnega načrtovanja, če hočemo družbeno neenakost v skupnosti še naprej vzpostavljati tudi z jezikom. To je zelo dobro delovalo v slovenskem prednarodnem in zgodnjenarodnem obdobju: dejavno je obvladala knjižni jezik (in tudi javno sporazumevanje) le peščica izobražencev.[20] Njihovo število je postopoma raslo, a so ostajali še dolgo v varni družbeni manjšini. Večini govorcev in govork slovenščine je bilo sprva namenjeno samo razumeti slovenščino knjižnih besedil (lahko so tudi občudovali lepoto jezika slišanega ali prebranega), ni pa bilo mišljeno, da bi jo tudi sami pisali ali govorili (Stabej 2010: 32). Z nadaljnjim razvojem obveznega splošnega izobraževanja je sicer sčasoma tudi za večino jezikovne skupnosti obveljalo, da se mora naučiti temeljev dejavne pisne rabe slovenščine – toda velika večina ni tega znanja nikoli uporabljala, tudi če je prišla do njega, saj je bil dostop do knjižne produkcije, pa tudi do javnega diskurza družbeno zelo omejen, saj je bila to zadeva izobražene manjšine. Kako se je v tem kontekstu načrtoval korpus slovenščine, ali bolj po domače, kako se je razvijal knjižni jezik? V zaprtem dvojnem krogu izobraženske elite:[21] a) elita je tvorila besedila; elita je pri sprejemanju besedil razsojala, kaj je jezikovno po njihovem okusu in volji in kaj ne in na podlagi tega revidirala svojo nadaljnjo jezikovno prakso; b) na podlagi besedil in jezikovnih stališč elite so nastajali knjižnojezikovni priročniki, slovarji in slovnice – in te je pri svoji nadaljnji besedilni produkciji v knjižnem jeziku uporabljala (če že) edinole elita. Večina govorcev in govork slovenščine je ostajala izključena iz tega dvojnega kroženja – razen tistih izjem, ki se jim je uspelo prebiti v elito.[22] Izključenost iz standardnojezikovnega kroga pretežnega dela večine dolgo časa ni čisto nič motila, saj je to sodilo v pričakovano podobo družbene urejenosti. Toda razmerja so se začela spreminjati in za današnji čas lahko domnevamo, da ostanki take dvopolnosti v jezikovni skupnosti pomenijo družbeno anomalijo, oziroma drugače: pomenijo nevarnost za nadaljnjo celovitost in povezanost jezikovne skupnosti. Če se danes večina govork in govorcev čuti simbolno in praktično izključena iz standardnojezikovnega kroga slovenščine (ker imajo občutek, da so kljub vsem dolgim letom šolanja slabo pismeni, da je njihov jezik slab in poln napak,

20 Pri razpravi o obvladanju knjižnega jezika konec 18. in dobri dve tretjini 19. stoletja moramo knjižni jezik razumeti skoraj v dobesednem pomenu: obvladovanje take slovenščine, da je dovolj dobra za v knjigo (ki je bila pred razmahom publicistike v slovenščini v 40-ih let 19. stoletja takorekoč edini pisni medij). Kako zelo raznolika pisna slovenščina je to bila (ne gre le za regionalne, temveč tudi za čisto individualne razlike), je splošni slovenski javnosti zelo malo znano, saj to v kontekstu stereotipne nacionalne ideologije kot presežena ovira na poti k enotnosti sodi v ropotarnico zgodovine in zato tudi ni vključeno v šolske programe. Zanimivo pa prav gotovo je – dober vir odkrivanja heterogenosti pisnih praks zgodovinske slovenščine je portal Jezikovni viri starejše slovenščine IMP (http://nl.ijs.si/imp, dostop 30. 7. 2015).

21 Elita je morda nekoliko neposrečen izraz, saj predstavnike elite pogosto miselno povezujemo tudi s premožnostjo, družbenimi privilegiji in podobnim; večina slovenskih izobražencev v narodni zgodovini s to predstavo nima prav veliko skupnega; da pa so izstopali po svojem družbenem položaju, vendarle drži.

22 Čisto izključeni niso bili; v 19. stoletju so neizobraženi govorci in govorke obveljali za dragocen vir čiste in prave slovenščine pri njeni standardizaciji, vendar pogojno in z omejitvami; za pridobivanje čiste slovenščine je bilo treba diskurzivno rudo teh govorcev prej temeljito očistiti jalovine ter tujih primesi, zato je bilo tako rudarjenje zahtevno in naporno. V 20. stoletju se je za tovrstno rudarjenje specializirala dialektologija; knjižnojezikovno slovaropisje pa se je temu viru s spremenjeno metodologijo (in najbrž tudi ideologijo) odpovedalo; SSKJ vsebuje narečne izraze samo, če jih je v svojih besedilih uporabljal kdo od elite (torej kdo med pisatelji in pesniki), in sicer v tistih besedilih, ki so vredna uslovarjenja.

da potrebujejo za vsako svoje pisno besedilo pomoč lektorja, da ne vejo, kaj je v jeziku prav in kaj narobe in kje bi prišli do informacij o tem ipd.) in hkrati onih, ki so del standardnojezikovnega kroga, ne dojemajo več kot družbeno upravičene in spoštovanja vredne elite, ki bi bila upravičena do kakršnega koli družbenega prestiža, imamo Slovenci in Slovenke težavo: to so namreč razmere za izstopanje iz jezikovne skupnosti.

6 SKLEP

Je torej treba slovensko jezikovno skupnost reševati – in je morda rešitev v popolnoma nasprotnem predlogu? V radikalnem načelu opisnosti in vsevključevalnosti, kot smo ga opredelili na koncu 3. poglavja? Kaj se zgodi, če se odločimo, da so vsi diskurzi vseh govorcev in govork skupnosti vredni uslovarjenja?[23] S tem popolnoma odpade tradicionalni koncept kultivacije in homogenizacije jezika, hkrati pa se jezikovna skupnost simbolno izhodiščno izenači, dehierarhizira. Taka radikalizacija diskurzivnega inputa jezikovnega opisa (seveda odvisno od tega, kako opis izpeljemo)[24] prinese s sabo ključno vprašanje – zakaj bi bila potem npr. slovar, pa tudi slovnica sploh še potrebna, če je vsa jezikovna praksa dobra prav taka, kot je? Se odgovor skriva v tem, da slovar postane pač evidentiranje vseh jezikovnih praks in je njegova morebitna usmerjevalna vloga pač izrazito drugotna in odvisna od tega, kaj z evidentiranim, opisanim naborom vseh jezikovnih praks pač nekdo hoče početi? Bi bil tak slovar torej po statusu podoben obsežni, če že ne izčrpni entomološki zbirki? Bi bil tak slovar priložnost za jezikovni voajerizem (poglejmo, kako to počnejo drugi), bi bil tak slovar pravzaprav neke vrste resničnostni šov?

Klasična jezikovnonačrtovalna predstava je, da jezikovno skupnost spreminjaš predvsem s tem, da spreminjaš jezikovno prakso govorcev in govork. Je cilj zares deskriptivnega sodobnega slovarja slovenščine morda to, da jezikovne prakse skupnosti ne spreminja? Oziroma da vsaj odveže jezikoslovje te naloge – saj se jezikovna skupnost in z njo njen jezik itak v vsakem primeru spreminjata v odvisnosti od kompleksnih družbenih okoliščin? Morda – a kot smo nakazali večkrat: kaže, da slovensko jezikovno skupnost še najzanesljivejše spremeniš, če ji ne usmerjaš in preoblikuješ jezika. Spremeniš jo na bolje. Naj Slovenec vidi Slovenca v slovarju, kakor vidi svoj obraz v ogledalu, če si sposodimo Levstika. Naj si narod končno privošči realnost, če si sposodimo Crnkoviča …

23 Vsaka totalnost se seveda slej ko prej sesede v partikularnost; čim rečemo skupnost, že vpeljemo ločevanje ljudi v skupnosti od ljudi zunaj nje; najmanjši skupni imenovalec ljudi v slovenski jezikovni skupnosti je seveda njihova jezikovna zmožnost v slovenščini; kaj pa je njen najmanjši možni obseg, ki posameznika ali posameznico že uvršča v skupnost, je seveda odprto vprašanje.

24 Če sicer v opisu zajameš vse diskurze, pa nekatere jezikovne pojave iz nekaterih diskurzov označiš za družbeno manj vredne, s tem dodatno utrjuješ vrednostno hierarhizacijo jezikovne skupnosti, morda celo bolj, kot če nekatere diskurze pri opisu prezreš in se z njimi sploh ne ukvarjaš.

Med ideologijo knjižnega in standardnega jezika

Vojko Gorjanc, Simon Krek in Damjan Popič

Abstract

This paper presents a sociolinguistic outline for developing a standard language corpus and the processes of standardisation, codification and modernisation; within this framework, the ideology of standard language is discussed. This framework serves as the basis for an examination of the normative context in Slovenian, especially in lexicographic terms. Standard-language linguistic environments such as Slovenian are characterised by various ideologies for *standard* or, in our case, *literary* language. Therefore, this paper focuses on presenting the ideology of literary language through the lens of lexicographic descriptions. Owing to the specific nature of the ideology, as well as the changes it has witnessed over the last five decades, we take issue with the term *literary language* and instead present the concept of *standard language*, which has a different ideological underpinning. With this we support the proposal for the consistent use of the term *Standard Slovenian* in contemporary linguistic descriptions of the Slovenian language.

Keywords: standardisation, codification, modernisation, literary language, standard language, ideology

Ključne besede: standardizacija, kodifikacija, modernizacija, knjižni jezik, standardni jezik, ideologija

1 UVOD

V prispevku na kratko predstavimo sociolingvistični okvir načrtovanja korpusa standardnega jezika, procesa standardizacije, kodifikacije in modernizacije. V ta kontekst v nadaljevanju umestimo razmišljanja o slovenskem jeziku, pri čemer problematiziramo uporabo termina *knjižni jezik*, in sicer tako zaradi sprememb v jezikovni skupnosti kot tudi sprememb v jezikoslovni teoriji in metodologiji ter tudi sprememb, ki jih je skozi čas doživel sam koncept *knjižnega jezika* v slovenskem prostoru. V nadaljevanju predstavimo koncept *standardnega jezika* ter argumentiramo predlog za njegovo dosledno rabo v sodobnih opisih slovenskega jezika, temelječih na načrtno zbranih podatkih o jezikovi rabi v korpusih slovenskega jezika. Pri jezikovnem opisu, ki temelji na korpusnih podatkih, namreč z dosledno uporabo korpusnojezikoslovnih metod lahko pridobimo podatke o jezikovnem standardu, tj. o jezikovni normi kot družbeno sprejeti lastnosti jezika; s tem pa se odrekamo opisu, ki vključuje jezikoslovno intervencijo v smislu jezikovnega kultiviranja, torej izključevanju tistih jezikovnih prvin, ki jih jezikovna skupnost uporablja, pa naj po takem ali drugačnem jezikoslovnem prepričanju ne bi bile primerne za sodobno pisno slovenščino. Če namreč v svoj koncept vključuje tudi tako načelo kultiviranja, torej jezikovno intervencijo s pozicije moči, združuje nezdružljivo, jezikovni opis na podlagi realnih podatkov o jeziku ter intervencijo v tako pridobljene jezikovne podatke v postopkih interpretacije teh podatkov. Tak opis ohranja in še poglablja dihotomijo med domnevno nekompetentnimi govorci oziroma pisci slovenskega jezika in pooblaščenimi razsodniki o jezikovni normi, s čimer se ne opolnomoči uporabnikov slovenskega jezika, ampak vzdržuje poseben status pooblaščenih inštitucij in posameznikov.[1]

Na ta način se programsko umeščamo v okvir idej poststrukturalnih kritičnih teorij, v okviru jezikoslovja predvsem kritične analize diskurza (Fairclough 2001; Dijk 2001) in kritične stilistike (Jeffries 2010; Jeffries in McIntyre 2010), ki v jezikoslovju naslavljajo izrazito družbena vprašanja in zagovarjajo angažirano pozicijo raziskovalca, njihov skupni imenovalec pa je humanistična želja po spremembi neenakosti v družbi in vzpostavljanje pogojev za enakopravni dialog posameznikov oziroma družbenih skupin. Pri tem so jezikovni opisi, ki imajo v svojem izhodišču uporabnika in skušajo razreševati uporabniške jezikovne težave, del tovrstnih prizadevanj, še posebej ko gre za jezikovno-kulturne prostore, zaznamovane s kulturo standardnega oz. v slovenskem primeru knjižnega jezika, kjer se v kulturni krog knjižnega jezika spuščajo ali iz njega izločajo posamezniki ali skupine na podlagi ocene o kompetentnosti knjižnojezikovnega govorca (Milroy 2001: 535–536). Hkrati z angažiranim pristopom pa se jasno zavedamo tudi

1 Glej tudi Odziv na objavo osnutka koncepta Novega slovarja slovenskega knjižnega jezika Centra za jezikovne vire in tehnologije UL (http://www.cjvt.si/projekti/sssj/odziv-na-objavo-osnutka-koncepta-nsskj/, dostop 4. 8. 2015). V prvem delu prispevka razpravljamo o sociolingvističnem okviru in vprašanjih ideologije na splošno, v nadaljevanju pa se usmerimo v slovarski kontekst.

pozicije raziskovalcev, ki delamo na področju humanističnih in družboslovnih ved in smo umeščeni v zgodovinske, družbene in kulturne kontekste ter je naše delovanje pogojeno z lastnimi prepričanji in ideološkimi pozicijami (Dijk 1993: 253; Katnić Bakaršić 2012: 7). Prepričani smo, da je tako izhodišče tisto, ki je prava pot k demokratizaciji tako diskurznih praks kot tudi znanstvenih diskurzov, saj je prav sklicevanje na ideološko in vrednostno nevtralnost v humanistiki in družboslovju prej znak za prikrito ideološkost in zavestno zavajajoče delovanje kot nevtralnost (Joseph in Taylor 1990: 2).

2 STANDARDIZACIJA IN STANDARDNOJEZIKOVNA KULTURA: OD KOD IDEOLOGIJA?

Sociolingvistični pojem jezikovne standardizacije je razmeroma univerzalen: jeziki gredo skozi faze izbire (izhodiščne jezikovne variante), kodifikacije, stabilizacije in vzdrževanja, ki vključuje nenehno modernizacijo (Cooper 1989: 31–32; 125), jezikovni standard pa naj bi potem veljal za vse udeležence javne komunikacije (Škiljan 1999: 168).

> V postopku standardizacije se predpiše izrazna podoba, optimizirajo slovnična sredstva ter fiksirajo pomeni besed. Glavni kodifikacijski priročniki so torej pravopis/pravorečje, slovnica in pravopisni slovar, ki določajo merila pravilnosti v okviru standardiziranih besedil. Produkt standardizacije je urejen jezik, očiščen nedoslednosti v pisavi/izreki, slovničnih dvoumnosti in redundanc ter motečih polisemij. /.../ Pojem standardnega jezika je povezan predvsem s tehničnimi vidiki jezikovne rabe: to je najprimernejša oziroma najbolj optimalna oblika jezika za pisavo, zlasti za tisk za množično komunikacijo, pa tudi za nekatere žanre, npr. znanstvena besedila ali administrativne obrazce, kjer mora biti jezik kar najbolj transparenten in objektiven (Skubic 2005: 46–47).

V okviru koncepta jezikovne standardizacije je potrebno zavedanje o tem, da je norma za vse pravzaprav nekaj povsem iluzornega (Cooper 1989: 134). V ozadju je velikokrat prav nasprotna logika, in sicer logika različnih družbenih elit s končnim namenom promocije svojega jezikoslovnega modela, ki vzpostavlja razlikovanje med partikularno elito in vsemi ostalimi (Cooper 1989: 135), pri čemer so bile zgodovinsko gledano družbene elite tipično povezane s pisnim jezikom kot prestižno obliko komunikacije (Cooper 1989: 137). Pisnost je bila torej v veliki meri tista, ki je bila medij in merilo družbene elite, kar pa se je z razvojem družb bistveno spremenilo. Če je pred stoletjem pisala in javno komunicirala le peščica družbeno privilegiranih posameznikov, je danes situacija popolnoma drugačna, pisno in javno komuniciramo praktično vsi pripadniki sodobnih družb, o čemer razpravljamo še v nadaljevanju.

Standardiziran jezik, ki je v določenih okoljih in obdobjih lahko ključen za skupno identifikacijo določene jezikovne skupnosti, lahko hkrati vzpostavlja tudi jezikovno situacijo, poznano kot »standardnojezikovna kultura« (Milroy 2001: 530, 535–536). V resnici pripadniki standardnojezikovne kulture zelo jasno definirajo svoj kulturni krog, ki v veliki meri temelji na preskriptivnih normativnih modelih, na jasnem ločevanju med pravilnim in napačnim, četudi na videz v popolnoma opisnem pristopu, preprosto zato, ker je v ozadju modela ideologija oz. niz prepričanj o jeziku, da je tudi pri jezikovnih vprašanjih treba upoštevati popolno skladnost z normo, to je jezikovno pravilnost (Milroy in Milroy 1999: 1). Standardizacija je torej tesno povezana z ideologijo standardnega jezika, prepričanjem, da obstaja pravilni način jezikovne rabe standardnega jezika, ki bi ga vsi pripadniki določene skupnosti morali uporabljati (Cooper 1989: 135), in torej temelji na nenehni dihotomiji standardno – nestandardno, pogojeni prav z ideologijo standardizacije in osredinjenosti na standardno varianto jezika (Milroy 2001: 534).

Razmerje med opisovanjem v jeziku in jezikovnim predpisovanjem še zdaleč ni enoznačno. Čeprav bi glede na naravo enega in drugega lahko pritrdili mnenju, da naj bi se jezikoslovje za preskripcijo zanimalo približno v tolikšni meri, kot se astronomija za astrologijo, pa po drugi strani drži, da preskriptivistična razpravljanja v jezikovni skupnosti in o jezikovni skupnosti odražajo odnos jezikovne skupnosti do jezika in na ta način lahko pozitivno vplivajo na kohezivnost in identiteto jezikovne skupnosti (Davies 1997: 4; Milroy in Milroy 1999: 3–4). S tega zornega kota je na eni strani opis jezika res ključna pot do jezikovnega standarda, hkrati pa preskripcija vpliva na občutek vključenosti v jezikovno skupnost, saj si pripadniki določene jezikovne skupnosti delimo normativne predstave o jeziku, predstave o jezikovnem obnašanju na različnih nivojih jezikovne rabe (Crystal 1987: 2; Davis 1997: 4–5).

Tovrsten odnos jezikovne skupnosti daje legitimacijo tako posameznikom kot inštitucijam, da določen jezikovni model promovirajo in s tem omogočajo vzdrževanje privilegiranih pozicij tistih, ki v določenem kontekstu o tem, kaj je jezikovni standard, odločajo, torej imajo nad jezikovnim standardom nadzor. V jezikovni skupnosti je vzpostavljeno pričakovanje po »vodenju s strani privilegiranih inštitucij« (Milroy 2001: 536), hkrati pa je inštitucijam v interesu, da se tak model ohranja pri življenju, kar pomeni, da nasprotujejo kakršnimkoli spremembam (Cooper 1989: 134, 135; Milroy in Milroy 1999: 4). Pri tovrstnem modelu delovanja je v odnosu do jezika izrazito opazna tendenca k preferiranju starejših jezikovnih oblik in pomenov ter želja po ohranjanju ali vzpostavljanju razlike med enim in drugim jezikovnim fenomenom, predvsem s stališča njegove standardnosti oz. nestandardnosti (Greenbaum 1988).

Za vsak standardni jezik je v njegovem razvoju ključna modernizacija, nenehna zmožnost prilagajanja novim komunikacijskim izzivom; ta pa je lahko izpeljana

na različne načine, skladno s celotno idejo posameznega kulturnega prostora in postopki jezikovne standardizacije v njem, (1) skozi model kultiviranja, ki vključuje tudi jezikovno intervencijo in velja za elitističnega (Haugen 1983) ali (2) z bolj demokratičnim pristopom, kjer se modernizacija razume kot amalgam novega in starega, glede na potrebe sociokulturnega prostora, pri čemer je jasno, da si ideje modernizacije v različnih kulturnih prostorih med posamezniki in skupinami nasprotujejo, nenehno pa tudi v tem okviru seveda delujejo centri moči, ki imajo v modernizacijskem procesu lahko bistveno lažjo nalogo vplivati na odločitve, kot to lahko počnejo posamezniki (Fishman 1989: 379–380). Poleg prave modernizacije pa je spreminjanje standarda povezano tudi z drugimi jezikovnimi intervencijami, npr. ideologijo politične korektnosti (Goddard in Patterson 2000: 73) ali ideologijo purizma (Thomas 1991); v teh primerih seveda ne gre za modernizacijo v pravem pomenu, saj se v jeziku ne pojavijo nove komunikacijske funkcije, gre le za spremembe oblike v okviru istih jezikovnih funkcij (Cooper 1989: 154).

Pomembno je poudariti, da se sodobne standardnojezikovne kulture v tako vzpostavljenem jezikovnem modelu težko znajdejo v demokratičnem procesu. Čeprav je v demokratičnih družbah nesprejemljiva diskriminacija glede na spol, socialno, nacionalno pripadnost, osebne okoliščine itd., pa standardnojezikovne kulture ohranjajo sprejemljivost jezikovne diskriminacije prav z vidika jezikovnega standarda (Milroy in Milroy 1999: 2–3), posledično v resnici tudi glede na regionalno, socialno, nacionalno itd. pripadnost, zato je naloga sodobnih demokratičnih družb premišljati tovrstna vprašanja in iskati rešitve, kako odgovoriti na tovrstne izzive v jezikovnih skupnostih – tudi ko gre za vprašanje ideologije standardnojezikovne kulture.

3 Slovenski knjižnojezikovni kontekst in slovar

Slovenska jezikovna situacija je klasična situacija standardnojezikovne kulture, le da je pri standardizaciji vlogo določitve kulturne skupnosti odigral knjižni jezik, torej moramo v slovenskem primeru govoriti o knjižnojezikovni kulturi. Čeprav je slovenistično jezikoslovje zavračalo izenačevanje norme in kodifikacije, je eksplicitno povezovanje norme s predpisovalno dejavnostjo prisotno,

> o čemer priča tudi SSKJ-jevsko izenačevanje norme s kodifikacijo (*norma = kar določa, kakšno sme, mora biti kako ravnanje, vedenje, mišljenje; pravilo, predpis*), tudi v terminološki zvezi norma knjižnega jezika (*jezikovna sredstva, možnosti, ki se smejo, morajo uporabljati v določenem knjižnem jeziku*) (H. Dobrovoljc 2015).

Kljub strokovnemu ločevanju jezikovne norme in kodifikacije (Müller 1982: 294–295; H. Dobrovoljc 2015), je iz zgornjega navedka razvidno, kako tesno je knjižni

jezik povezan s kodifikacijo; v razmislekih o knjižnem jeziku ni torej prisotna le ideja jezikovnega opisovanja, ampak hkratnega reguliranja knjižnega jezika, kaj se sme oz. mora uporabljati.[2] Prav na ta način, torej z merili jezikovne pravilnosti, se najlažje določa okvire knjižnojezikovne kulture oz. knjižnojezikovne skupnosti, torej prepoznava, kdo v to skupnost sodi in kdo ne. Nas pa v nadaljevanju normativni in knjižnojezikovni koncept zanima v okviru slovarske dejavnosti.

Sodobno slovensko slovaropisje sega v konec 19. stoletja, eden od mejnikov v standardizaciji slovenskega jezika pa je Pleteršnikov Slovensko-nemški slovar (1894–95). Zaradi umestitve v okolje, ki je v resnici veliko bolj kot enojezični slovarski opis potrebovalo dvojezičnega, je norma slovenskega jezika sopostavljena nemški, a hkrati slovar

> oblikuje svojo vrednotenjsko (normativno) zasnovo na lastnostih slovenskega jezika, izpričanih na bogatem gradivu. Prav z gradivno členjenostjo slovenskega besedišča glede na njegove socialne zvrsti – podatki o narečnosti, zbornosti (vsaj deloma oznaka novo knjižno) – glede na kronološko zaznamovanost – podatki o času rabe oziroma avtorju – do neke mere tudi o slogovnih posebnostih, priča o jezikovni ustaljenosti in zato tudi lastni normativni zmogljivosti slovenščine (Vidovič Muha 1992: 10).

Pleteršnik je s svojim delom vzpostavil okvir objektivnega pogleda na slovarski opis, saj temelji na predstavitvi aktualnega dokumentiranega jezikovnega gradiva (Vidovič Muha 2013: 256). Zanimivo je, da je slovar z vidika normativnosti v slovenskem prostoru precej različno vrednoten, od ocene o njegovi moderni normativni podobi, temelječi na sodobni metodologiji (Vidovič Muha 1992: 10), do ocene o normativni omahljivosti (H. Dobrovoljc 2004: 46), saj naj bi zaradi svoje variantnosti ne prispeval k poenotenju knjižne norme (Majcenovič 1999: 87), kar kaže na to, da standardizacijski postopek tudi zgodovinskorazvojno opazujemo z vidika nevariantosti, s čimer je tak pogled na normo osredotočen na jezikovno uniformnost in učinkovitost, torej je v resnici bližje normativnosti, kakršno poznamo npr. pri merskih enotah, kot pa normi, izhajajoči iz realnosti komunikacije, ki teži k variantnosti (Cooper 1989: 133).

V različnih obdobjih 20. stoletja je slovensko normativistiko v veliki meri zaznamoval različen odnos do jezikovne variantosti ter jezikovne pravilnosti; pri slednjem tudi v obliki paranormativnih priročnikov, ki so razmerje do jezikovne pravilnosti zaostrovali predvsem na podlagi ideologije purizma (H. Dobrovoljc 2004: 55–56). Ves čas pa se zdi, da je bila ključna vloga knjižnega jezika povezovalna, še posebej takrat, ko se je v javnosti obudilo ali izostrilo vprašanje o ogroženosti jezika; takrat so prav jezikovna vprašanja – in to prav tistega tipa, o

2 Definicija termina *norma knjižnega jezika* v SSKJ2 ni spremenjena. http://www.sskj2.si/iskanje?Mode=Headword&Query=norma (dostop 4. 8. 2015).

katerih smo razpravljali kot o tipičnih vprašanjih knjižnojezikovne kulture, torej vprašanja jezikovne pravilnosti – posredno služila tudi za mobilizacijo narodne zavesti, kot je bilo to npr. pri Jezikovnem razsodišču (Kmecl 2005: 87).

Bistveno spremembo normativnega pogleda v slovenski prostor prinesejo sodobna spoznanja o socialno- in funkcijskozvrstni členjenosti jezika predvsem v 60. letih 20. stoletja. V tem duhu je nastajal tudi koncept slovarja slovenskega jezika, temeljnega normotvornega priročnika, ki je bil do takrat nadomeščan z zasilnimi rešitvami, predvsem pravopisnimi slovarji, s slovarsko zasnovo pa dobimo tudi jasno definiran koncept *slovenskega knjižnega jezika*.

> Slovenci smo navajeni, morda bolj kakor drugi narodi, da zaradi narodnostne ogroženosti, da se v knjižni jezik ne vnaša preveč tujega, oz. tega, česar ne izkazuje literarna tradicija. Zdaj bo v slovarju registriranega mnogo več: to, kar je bilo priznano kot dobro, manj dobro in tudi to, kar je veljalo za slabo. Hoteli smo prikazati knjižni jezik v najširšem pomenu besede: živ, poln, z dubletami, notranjimi nasprotji, vzporednimi istočasnimi normami, jezik sredi zagona in razvoja. /.../ Slovar bo registriral dejansko stanje v jeziku, torej osnove njegove norme, s kvalifikatorji in kvalifikatorskimi pojasnili pa bodo vstavljene v ta okvir posebnosti, dvojnosti in izjeme (Suhadolnik 1968: 4–5).

Uvod v prvo knjigo Slovarja slovenskega knjižnega jezika (1970; dalje SSKJ) opredeljuje vsebino slovarja eksplicitno kot slovar slovenskega knjižnega jezika:

> V slovarju je zajet besedni zaklad (besede, zveze) in prikazana njegova raba, kakor se kaže **v sodobnem slovenskem knjižnem jeziku**, to je v obdobju **od začetka tega stoletja do 1969** oziroma do leta izida posamezne knjige. Obsega vse **bistvene prvine knjižnega jezika**: leposlovni, znanstveni, publicistični, časopisni, pogovorni jezik, terminologijo, žargone in narečno besedišče. (Poudarili avtorji.)

V slovarju se je udejanjil koncept slovenskega knjižnega jezika »v najširšem pomenu besede: živ, poln, z dubletami, notranjimi nasprotji, vzporednimi istočasnimi normami« (Suhadolnik 1968: 4), torej knjižni jezik »kot skupek vzporedno živečih, enakovrednih stilov«, pri čemer »noben stil ni kriterij za ocenjevanje drugega«, za vse pa velja, »da gre pri vseh, v slovarju potencialno predstavljenih stilnih realizacijah knjižnega jezika, za enoten, trden knjižni sistem« (Vidovič Muha 1972: 178).

Taka zasnova se v slovarskih opisih začne rušiti s Slovarjem novejšega besedja slovenskega jezika (2012; dalje SNB), ki se samoopredeljuje kot dodatek k SSKJ in »/p/rinaša besede, ki se na novo uporabljajo ali na novo uveljavljajo v slovenskem pisnem jeziku, to pa je danes precej širši pojem kot knjižni jezik« (Snoj 2013).[3]

3 http://www.mladina.si/120969/dr-marko-snoj-danes-stare-mame-govorijo-ful/ (dostop 31. 7. 2015).

Jasna konceptualna zasnova v SSKJ tu trči ob popolnoma drugačno idejo normativnosti in knjižnosti; slovarju se celo odreka normativnost, ki se pripisuje zgolj pravopisu:

> Veliko teh besed je tako rekoč knjižnih. Je pa res, da pojma knjižni jezik danes ne uporabljamo prav pogosto. Govorimo raje o standardnem jeziku, čeprav tega nihče zares natančno ne definira. Sam sicer menim, da sta standardni in knjižni jezik pravzaprav sinonima in da je današnja uporaba prilastka standardni namesto knjižni prevladala pod vplivom angleščine. /.../ Nekatere tu prikazane besede verjetno ne bodo nikdar standardne, kaj šele kanonizirane kot norma. Naš osnovni normativni priročnik tako ostaja Slovenski pravopis 2001 (Snoj 2013).

Tako je glede normativnosti predvidljiv konflikt med zasnovo SSKJ in SSKJ2 (2014): kako sicer koherenten normativni koncept, a vendarle koncept izpred 50 let, aplicirati na slovar, ki vsebuje tudi sodobno ali sodobnejšo leksiko? V kvalifikatorski sistem »so redaktorji SSKJ2 dokaj močno posegli in to je pravzaprav ena od večjih težav končnega izdelka«:

> Splošni vtis pri prilagajanju kvalifikatorskega sistema je ta, da je šlo veliko prizadevanja v spremembe izbrane peščice kvalifikatorjev po celotnem slovarju, vendar je rezultat brez podrobnega razumevanja sistema v SSKJ1 nenavadna mešanica »stare« in »nove« slovenščine, pri kateri je zdaj težko razumeti, kaj je sodobno in kaj dejansko spada ali je spadalo v določene zvrsti jezika nekoč in danes (Krek 2014: 145).

Z načinom prenove prve izdaje SSKJ in pristopom do normativnosti v SSKJ 2, kot se ta izraža skozi rekvalifikacijo in presistematizacijo kvalifikatorskega sistema, imajo težavo tudi avtorji prve izdaje:

> Slovenski jezik ima ustaljene in normativno preizkušene smernice glede sprejemanja prevzetih leksikalnih enot v pisavi in izgovoru, ki jeziku na izrazni ravni zagotavljajo ustrezno kultiviranost. Druga izdaja SSKJ teh smernic ne upošteva in se brez distance podreja številčnemu diktatu korpusov, namesto da bi v skladu s konceptom SSKJ upoštevala tradicijo kultiviranja slovenskega jezika, kot jo poleg načel Slovenskega pravopisa neposredno potrjujejo tudi slovenska besedila zborne zvrsti (Ahlin et al. 2014: 124).

Odlomek lepo izpostavi temeljno težavo SSKJ2, vprašanje kultiviranja slovenskega jezika, torej postopka, ki je v modernizaciji jezika glede na koncept knjižnega jezika nujni element jezikovnega opisa, če naj bo ta opis *slovenskega knjižnega jezika*.

Prva izdaja SSKJ in SNB sta bila združena v drugi izdaji SSKJ, kjer se že v opredelitvi slovarja razgali težava, na katero nakazuje razumevanje normativnosti v SNB:

SSKJ
V slovarju je zajet besedni zaklad (besede, zveze) in prikazana njegova raba, kakor se kaže v sodobnem slovenskem knjižnem jeziku, to je v obdobju od začetka tega stoletja do 1969 oziroma do leta izida posamezne knjige. Obsega vse bistvene prvine knjižnega jezika: leposlovni, znanstveni, publicistični, časopisni, pogovorni jezik, terminologijo, žargone in narečno besedišče.

SSKJ2
V slovarju je zajeto besedje (besede, zveze) in prikazana raba slovenskega jezika, kakor se kaže predvsem v zapisanih besedilih v obdobju od druge polovice 19. stoletja do 2013. Obsega vse bistvene prvine slovenskega jezika: v okviru socialnih zvrsti tako knjižne kot tudi neknjižne zvrsti in interesne govorice oziroma od praktičnosporazumevalnega prek strokovnega jezika in publicistike do umetnostnega jezika v okviru funkcijskih zvrsti.

V SSKJ je knjižni jezik dojet kot jasno opredeljen in omejen nabor stilov (leposlovni, znanstveni, publicistični, časopisni, pogovorni jezik, terminologija, žargoni in narečno besedišče), ki je »skupek vzporedno živečih, enakovrednih stilov« (Vidovič Muha 1972: 178). Vse, kar najdemo v slovarju, je del »knjižnega jezika«, ne glede na zvrstno opredelitev. Nasprotno pa se »sodobni slovenski knjižni jezik« iz SSKJ v drugi izdaji spremeni v »zapisana besedila« oz. »slovenski jezik« brez opredelitve knjižnosti, pri čemer so po novem v skoraj identičen slovar (prim. Krek 2014) vključene »tako knjižne kot neknjižne zvrsti«, z nedefiniranim konglomeratom socialnih in funkcijskih zvrsti, o katerih knjižnosti ali neknjižnosti ne vemo dosti. Celota vsebine SSKJ2 torej ni več »slovenski knjižni jezik«, tako kot v SSKJ, temveč je to »slovar slovenskega tako knjižnega kot neknjižnega jezika«, sami pa moramo v njem razločiti »knjižni« del od »neknjižnega«. Če je torej bralec pogledal v SSKJ in v njem našel iztočnico, je lahko predpostavljal, da je ta del slovenskega knjižnega jezika, z morebitnim opozorilom, v katero zvrst knjižnega jezika spada, in z opozorilom, v kakšnih okoliščinah je raba izraza primerna oz. kako bo v knjižnem jeziku ta beseda funkcionirala. Ta sistem je bil s subtilnim premikom postavljen na glavo in po novem gre za slovar *pisnega jezika* z drugačnim konceptom. Značilno je, da so to opazili le avtorji SSKJ, ki dobro vedo, kako je prva izdaja nastala:

> Nove leksikalne enote so besedilu prve izdaje mehanično dodane ne glede na koncept SSKJ, brez upoštevanja specifičnosti korpusnega gradiva, zlasti pa brez razvidnih meril o sprejemanju. Posledica tega je, da je med temi enotami množica besed, katerih sprejem je v nasprotju s konceptom SSKJ, npr. *džankizacija*, *džek*, *fopati*, *igličar* 'narkoman', *kulica*, *narkič* 'narkoman', *profi* 'profesionalec', *sfaliti*, *štekati*, *radodajka*, *lapati* itd. - Poudarjamo, da se gornja ugotovitev nanaša na neskladje s konceptom SSKJ, ki je slovar slovenskega knjižnega jezika, in ne na zavračanje možnosti, da so vse te leksikalne enote sicer sprejete v kateri drugi slovar slovenskega jezika, ki bi to predvideval v svojem konceptu (Ahlin et al. 2014: 122–123).

S tega izhodišča si oglejmo še napoved vsebine iz Osnutka koncepta novega razlagalnega slovarja slovenskega knjižnega jezika (NSSKJ), ki se še vedno opredeljuje za slovar »knjižnega jezika«:

> Knjižni jezik razumemo kot **uzaveščeni, kultivirani in konvencionalni nadregionalni kod**, ki ga v slovenskem jezikovnem okolju govorci slovenščine nezaznamovano rabijo zlasti v javnih in formalnih govornih položajih. Pri tem je **kultiviranost knjižnega jezika** razumljena kot **upoštevanje splošno veljavnih načel, norm in družbenih konvencij**. Knjižni jezik temelji na jezikovni tradiciji in se tudi s **kodifikacijo** utrjuje, v svojem jedru je razmeroma stabilen, čeprav se skozi čas postopoma spreminja. Njegova poglavitna vloga je, da kot **standardni kod** uresničuje osrednjo in temeljno sporazumevalno in povezovalno vlogo jezika za javno rabo. (Poudarili avtorji.)

Knjižni jezik tudi v tem slovarju ne bo več »skupek vzporedno živečih, enakovrednih stilov« (Vidovič Muha 1972: 178), ki so na enoten način opredeljeni v slovarju, temveč bo »kod«, ki se utrjuje s »kodifikacijo«, torej pravopisom in pravorečjem, ena njegovih glavnih opredelitev pa je »kultiviranost«. Skratka, slovar opisuje oboje: tako kultivirani kod kot nekultivirani preostanek, za zadnjega pa tako kot v primeru druge izdaje SSKJ vsaj na podlagi povedanega ne moremo biti prepričani, kaj vse zajema. Kvalificiranje socialnih zvrsti je v NSSKJ opredeljeno takole:

> Okvalificiranost v okviru socialne zvrstnosti določa odstopanje od **kultivirane**, v (javnih, formalnih) govornih položajih **ustaljene, predvidljive** in vsaj pretežno **s kodifikacijo usklajen**e leksike slovenskega knjižnega jezika, ki je večinoma razširjena tako v govorjenih kot zapisanih besedilih. Leksika, ki je torej **splošno sprejeta, ustaljena**, ima **uzaveščeno razmerje do norme** in je **kultivirana, primerna za formalne govorne položaje**, je torej knjižna in z vidika socialne zvrstnosti nezaznamovana. Zaznamovanost gre v dve smeri, bodisi proti izrazito omejeni rabi v posebej izbranih govornih položajih za izrecno poudarjanje jezikovne kultiviranosti in umetelnosti v smislu izkazovanja lastne jezikovne zmožnosti (tj. v ozkoknjižno) bodisi proti manj formalni rabi, tj. najprej v pogovornost (knjižno in neknjižno), po potrebi tudi v narečnost, sleng, otroški govor. (Poudarili avtorji.)

Načrtovani slovar vendarle je normativni, torej ne drži več teza, da je normativnost v slovarju rezervirana za pravopisni slovar, »knjižna leksika« (kultivirana, ustaljena, predvidljiva, splošno sprejeta, s kodifikacijo usklajena) v njem je nezaznamovana in ni kvalificirana. Če je zaznamovana, načeloma ni več knjižna. Kljub temu kvalifikatorja »ozkoknjižno« in »knjižno pogovorno« dajeta vedeti, da je tako opredeljena leksika še vedno del knjižnega jezika. Zelo očitno ni več knjižna leksika tista, ki je označena z »neknjižno pogovorno«, ni pa jasno, kaj je z

narečnim besediščem, slengom, otroškim govorom, še manj, kaj je s pripadnostjo knjižnemu jeziku pri funkcijskozvrstnih ali ekspresivnih kvalifikatorjih.

4 ZAKAJ NOVO IDEOLOŠKO IZHODIŠČE IN *SLOVENSKI STANDARDNI JEZIK*?

Odločitev za dosledno uporabo termina standardni jezik utemeljujemo tako z zunanjezikovnimi dejavniki kot tudi spremembami, ki smo jih glede prevrednotenja koncepta knjižnega jezika s strani privilegirane inštitucije v okviru »knjižnojezikovne situacije« v slovenskem prostoru opisali zgoraj. V bistvu gre za težnjo po demokratizaciji in hkrati vračanje na osnovna demokratična izhodišča, ki so bila po našem mnenju znotraj slovenskega prostora že prisotna, a v zadnjem času pozabljena ali zavestno zanikana. Od novih premislekov, kateri diskurzi so tisti, ki so relevantni za standardnojezikovni opis (o tem več v prispevku M. Stabeja 2015), do metodologije jezikovne analize in predstavitve jezikovnih opisov na način, ki bo izrazito uporabniško usmerjen (o tem v poglavju Slovarski uporabniki). Vse tudi z željo, da se prevrednoti koncept standardnojezikovne kulture: koncept jezikovne norme kot družbeno sprejete lastnosti jezika namreč omogoča njeno razširitev, jo odvezuje naveze na pooblaščene posameznike in inštitucije, hkrati pa zavezuje k večji odgovornosti glede jezikovne rabe oz. ji sploh omogoča, da odgovornost prevzame sama.

4.1 Nove okoliščine

Pojem knjižnega jezika je bil skozi zgodovino močno zaznamovan s svojo narodnozdruževalno in narodnopredstavno vlogo (Toporišič 1991: 392). Ta vloga je bila v preteklosti seveda pomembna, v trenutnih okoliščinah, ko lahko rečemo, da so razmere za slovenski jezik izrazito ugodne, in ko slovenščina – tudi zaradi ustreznega jezikovnonačrtovalskega delovanja – uživa prestižni položaj uradnega jezika EU (Stabej 2010: 191; Popič in Gorjanc 2014), lahko rečemo, da so za Slovenijo in slovenščino nastopile bistveno drugačne okoliščine. Osredotočili se bomo le na tiste, ki po našem mnenju terjajo razmislek, ko govorimo o standardizaciji danes.

Prva sta dva formalna dogodka, ki pomenita pomembno statusno spremembo in imata tudi zelo praktične posledice. Prvi je nastanek samostojne države l. 1991, s čimer je slovenščina postala uradni oz. državni jezik državne skupnosti, drugi pa vstop v Evropsko unijo l. 2004, s čimer je postala eden od uradnih jezikov te skupnosti. Ali z besedami iz osnutka Nacionalnega programa za jezikovno politiko 2012–2016: »Slovenščina je z vključitvijo med uradne jezike Evropske unije

pridobila večjo mednarodno težo na simbolni ravni, predvsem pa veliko operativnih možnosti za sodelovanje pri raziskovanju in uporabi jezikov v skupnosti s 23 uradnimi jeziki.«[4] To okoliščino poudarjamo predvsem zaradi (a) potencialne razbremenitve občutka ogroženosti, ki je bil prisoten v vseh večnacionalnih državnih skupnostih do ustanovitve samostojne države in je, kot smo videli, pomenil mobilizacijo jezikovne skupnosti tudi skozi normativna vprašanja jezika, ter (b) zahtev in pričakovanj, ki jih s seboj prinese status uradnega jezika EU – te so povezane predvsem z ustreznimi vzdrževalnimi in modernizacijskimi dejavnostmi. Ob tem pa je postala slovenska družba tudi veliko bolj odprta, povečane migracije na globalni ravni so vplivale tudi na slovenski prostor in slovensko jezikovno skupnost (Stabej 2010: 75). Slednje je pomembno tudi z vidika diskurznih praks, saj je logično, da tako spremenjene okoliščine vplivajo na komunikacijsko realnost, ki jo moramo tudi v procesih standardizacije premisliti na novo v razmerju do premislekov, ki so bili povezani z idejo bolj ali manj homogene nacionalne skupnosti.

Pomembna okoliščina, ki je močno spremenila jezikovno realnost, je bliskovit razvoj sodobnih informacijsko-komunikacijskih tehnologij, katerega posledice so vidne v delovanju celotne družbe, logično tudi v jezikovni rabi. Kot pravi Crystal (2011: 76), »nove tehnologije vedno povečajo slogovni razpon jezika« – tako se s tiskom poveča potreba po standardizaciji pisnega jezika, nastane vrsta novih formatov in uredniških standardov, z radiem in TV se poveča potreba po definiranju enotnega govorjenega standarda, s spletom in digitalizacijo se poveča predvsem možnost dvostranske ali večstranske komunikacije, ki je hkrati tudi javna. To pomeni, da je imel svetovni splet in splošni prehod s papirja v digitalno obliko daljnosežne posledice za proces javne objave in dostopa do besedil, saj se je število piscev z možnostjo javnega objavljanja dramatično povečalo, čas od nastanka besedila do javne objave se je bistveno skrajšal, veliko prej zasebnih žanrov pa je prešlo v javno sfero (forumi, klepetalnice, družabna omrežja itd.). Na drugi strani se je vloga (literarnega, časopisnega in revijalnega) založništva kot filtra med avtorjem in bralcem bistveno spremenila. Proces objavljanja se je individualiziral, klasični proces objavljanja v novih razmerah ne deluje več. Standardizacijski koncept za 21. stoletje bo moral upoštevati kvantni preskok pri množičnosti javne komunikacije, novo digitalno besedilno stvarnost, neomejen dostop do digitalnih (jezikovnih) virov in možnosti njihovega povezovanja, predvsem pa tudi vse možnosti uporabe napovedujočih se novih jezikovnih tehnologij.

Končno se temu pridružuje še razmislek, vezan na informatizacijo in globalizacijo koncepta znanja, ki sega od prej omenjene revolucije v informacijsko-komunikacijskih tehnologijah, vključno s semantičnim spletom (angl. *Semantic Web*), povezanimi podatki (angl. *Linked Data*), sistemi za organizacijo znanja (angl. *KOS*

4 http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Zakonodaja/2013/Resolucija_-_sprejeto_besedilo__15.7.2013_.pdf (dostop 30. 7. 2015).

– *Knowledge Organization System*), preko množičnih podatkov (angl. *Big Data*), ki v primeru tekstovnih podatkov po nujnosti vključuje (naravni) jezik, v katerem so ti podatki, vse do umetne inteligence. Če želimo, da bo nek jezik deloval v okoliščinah, v katerih je treba znati strojno procesirati vse jezikovne modalitete (pisni, govorjeni) in zvrsti (znanstveni diskurz, specializirani diskurz posameznih strok, od pravnega jezika do športa, financ, medicine, jezik socialnih omrežij, mladostniški sleng itd.), predvsem pa je iz teh informacij treba znati ustvarjati (novo) znanje, je treba tudi konceptualno na novo premisliti obstoječe pojavne oblike (slovenskega) jezika in iz tega izhodišča oblikovati pojmovnik različnih – recimo temu nevtralno – tipov jezika, vključno s standardizacijsko lestvico, ki upošteva zatečeno knjižnojezikovno slovensko situacijo. Prej omenjeno znanje seveda ni omejeno na slovenski jezik, temveč se kodirano v različne digitalne entitete (Wikipedija, Europeana, Google Knowledge Vault itd.) povezuje v ogromen globalni konglomerat procesljive vednosti, ki ga izkoriščajo umetnointeligenčni sistemi, kot so Watson podjetja IBM ali Wolfram Alpha. Ne biti znotraj teh procesov ustvarjanja globalne vednosti in seveda tudi njihovega izkoriščanja za vsak jezik pomeni, da bo kritično hendikepiran, po pomembnosti pa je problem mogoče primerjati z neugodno situacijo jezika brez državne skupnosti v času nastajanja nacionalnih držav.

4.2 Sprememba v konceptu opisa: nova teorija in nova metodologija

Velik del evropskega jezikoslovja še vedno temelji na strukturalistični tradiciji, v jezikoslovno raziskovanje pa niso bili vključeni poststrukturalni pristopi, evropska jezikoslovna misel zaostaja za sodobnim humanističnim in družboslovnim raziskovanjem (Motschenbacher 2010: 5). To še kako velja tudi za slovenistično jezikoslovje, ki le s težavo vključuje poststrukturalno jezikoslovje, še težje pa ga povezuje s poststrukturalnimi družboslovnimi raziskavami. Poststrukturalni jezikoslovni pristopi pri jezikoslovnem delovanju izpostavljajo razmerja opis – interpretacija – pojasnitev, s fokusom na slednjem (Fairclough 2001; Jeffries 2010). Opis ni več ultimativno končno jezikoslovčevo dejanje, ampak potrebuje nadgradnjo; pravzaprav se težišče delovanja premakne na pojasnjevanje. S tem pa vstopamo v razmerje do uporabnika, torej razmišljamo, kako jezikovne podatke interpretirati in jih uporabniku čim bolj učinkovito razložiti.

Čeprav se zdi, da naslavljamo nekatera vprašanja v konceptu opisa popolnoma na novo, jih le na novo premišljamo v okviru novejših poststrukturalnih pristopov. V resnici se v marsičem vračamo k izhodiščem sociolingvistike iz 60. in 70. let prejšnjega stoletja, ki pa jih slovenski prostor – predvsem zaradi prevladujoče strukturalistične misli – ni ne poznal, večinoma pa tudi ne jemal dovolj resno. Vprašanja o demokratizaciji, ko gre za načrtovanje korpusa (Neustupný 1968),

jezikovne norme glede na pozicije družbene moči in identitet v jezikovni skupnosti (Haugen 1968) ali jezikovnega razvoja in modernizacije korpusa (Fishman 1968), so že takrat oblikovala močno teoretsko podlago poststrukturalnega pristopa v jezikoslovju (Neustupný 1978).

Nov pa je vsaj od začetka 90. let 20. stoletja korpusnojezikoslovni teoretsko-metodološki pristop, ki v analizo jezika vnaša večjo verodostojnost in objektivizacijo raziskovanja – velik obseg načrtno zbranega gradiva namreč omogoča izpostavitev v jeziku tipičnega in zmanjšuje možnost interpretiranja le obrobnega kot temeljnega (Čermák 1995: 119). A to še zdaleč ni vse: sprememba ni le gradivna, ampak tudi teoretska in metodološka (Verdonik 2015). Prav zato, ker sta tako teorija kot metodologija v korpusnem jezikoslovju pomembna kot novo jezikoslovno izhodišče, sta se uveljavili poimenovanji korpus kot teorija in korpus kot metoda (McEnery in Hardie 2012: 150–151). Jezikovni opisi danes temeljijo na korpusnih podatkih, za njihovo relevantnost pa je pomembno dosledno upoštevanje tako korpusne teorije kot metodologije, sicer v opisu nismo naredili bistvenega premika glede na predkorpusne jezikovne podatke.

V zadnjih letih lahko spremljamo tudi intenzivno oblikovanje metod nove leksikografije ali natančneje e-leksikografije,[5] ki izhajajo iz korpusnega jezikoslovja, računalniškega jezikoslovja in – posledično širše – iz informatike. Pri tem ne gre za prestavitev konceptov klasične leksikografije v digitalno okolje, npr. elektronskega korpusa kot slovarskega gradiva v digitalni obliki, temveč za menjavo zornega kota glede pridobivanja, interpretiranja in predstavitve jezikovnih podatkov. E-leksikografija pri tem izkorišča metode, ki so bile v zadnjih petindvajsetih letih razvite v okviru strojnega procesiranja naravnih jezikov, pri čemer se fokusira na detektiranje in luščenje informacij o tistih jezikovnih pojavih, ki so zanimivi za vključitev v slovarski ali leksikalni tip podatkovnih baz. Te metode vključujejo vse jezikovne ravnine, od oblikoslovja (npr. prepoznavanje in tvorjenje oblikoslovnih vzorcev), leksikogramatike (npr. prepoznavanje večbesednih enot od kolokacij do frazeologije, luščenje vezljivostnih vzorcev), semantike (strojno prepoznavanje semantične podobnosti (angl. *semantic similarity*), avtomatsko detektiranje neologizmov, strojno prepoznavanje definicij itd.), fokus pa se prestavlja z obravnave statičnih korpusov na dinamično procesiranje jezika v času (množični podatki), z avtomatskim opozarjanjem na spremembe na točkah, kamor so postavljeni jezikovni senzorji.

Če je jezikoslovje od časa nastanka in prevlade strukturalnega jezikoslovja prešlo več faz, ki so bile v slovenskem prostoru šibko zaznane, se je leksikografija kot praktična dejavnost – tako kot mnoge druge, ki so bile radikalno transformirane z razvojem

5 Organiziranost področja e-leksikografije in njigovo dinamičnost izkazuje tudi sklop konferenc, od leta 2009 organiziranih na to temo, *eLex: Electronic lexicography in the 21st century* (https://elex.link/, dostop 3. 8. 2015), in ena od EU akcij COST *European Network of e-Lexicography* (http://www.elexicography.eu/, dostop 3. 8. 2015).

računalništva in informacijsko-komunikacijskih tehnologij – že pomaknila z jezikoslovne stroke na vmesno pozicijo med jezikoslovjem in informatiko, kar pomeni, da enakovredno črpa teorije in metode z obeh strani interdisciplinarnega razpona.

4.3 Sprememba v samem ideološkem konceptu knjižnega jezika

Kot smo lahko videli pri primerjavi med koncepti knjižnega jezika v SSKJ, SNB, SSKJ2 in NSSKJ, se je konceptualni model knjižnega jezika močno spremenil, iz koherentnega, ki je v slovenskem prostoru v veliki meri hkrati pomenil prelom s tradicionalno preskripcijo. V 60. letih jasno postavljen model, ki je odgovarjal tako takratnim teoretskim premislekom pri jezikovnem opisovanju kot družbenim okoliščinam, v katerih je nastal, se je spremenil do te mere, da ni več prepoznaven, hkrati pa tudi ni več konceptualno jasen.

Hkrati pa lahko tudi v jezikoslovnem diskurzu opažamo, da se je dojemanje knjižnega jezika močno radikaliziralo, v resnici do te mere, da je eksplicitno nacionalistično, s čimer se knjižni jezik še bolj elitizira in oži svoj kulturni, v resnici kulturno-nacionalni krog:

> Vsak državljan Republike Slovenije bi moral biti tudi govorec slovenskega knjižnega jezika, ki je jezikovna zvrst za izražanje najzahtevnejših vsebin v slovenskem jeziku in imajo tudi uradni značaj. Šele na ta način se lahko tudi celovito vključi v na jeziku temelječo slovensko družbo (Tivadar in Tivadar 2015: 43).

Ob takem zapisu je treba spomniti tako na 11. kot na 14. člen slovenske Ustave.[6] Prvi govori o tem, da je uradni jezik v Sloveniji slovenščina (torej slovenščina kot taka, ne zgolj knjižna slovenščina), drugi pa o tem, da smo državljani enakopravni,[7] ne glede na »katerokoli osebno okoliščino«, torej tudi to, ali smo govorci slovenskega knjižnega jezika ali ne oz. kakšno »slovenščino« govorimo. Ne glede na morebitno dobronamernost piscev je treba izpostaviti že latentno možnost diskriminacije državljanov RS na podlagi jezika, ki je implicirana v navedenem odlomku, ter to, da je ta možnost očitno konceptualno vpisana predvsem v pojem »knjižnega jezika«.

Tako izključujoče povezovanje slovenske družbe in njene jezikovne skupnosti s knjižnim jezikom v resnici le zapira krog njenih uporabnikov, s čimer se

6 https://zakonodaja.com/ustava/urs (dostop 3. 8. 2015).

7 »V Sloveniji so vsakomur zagotovljene enake človekove pravice in temeljne svoboščine, ne glede na narodnost, raso, spol, jezik, vero, politično ali drugo prepričanje, gmotno stanje, rojstvo, izobrazbo, družbeni položaj, invalidnost ali katerokoli drugo osebno okoliščino.« https://zakonodaja.com/ustava/urs/14-clen-enakost-pred-zakonom (dostop 3. 8. 2015).

identifikacijski krog posameznikov in družbenih skupin s tako elitističnim pogledom na knjižni jezik lahko posledično le oži in pomeni osnovo ne le za prestop v drug neknjižni model, ampak tudi izstop iz jezikovnega modela, torej tudi iz jezikovne skupnosti (prim. tudi Stabej 2015).

5 SKLEP

Za sodobni globalizirani funkcionalistično usmerjeni svet, v katerem se znanje in vednost bliskovito pretakata po spletu v vseh digitalno preživelih jezikih in v katerem sta večjezičnost ter čezjezičnost samoumevni in nista več vezani le na človeka, temveč tudi na računalniškega »uporabnika«, na novo radikalizirani koncept knjižnega jezika s centralno inštitucijo, ki presoja o jezikovnih pojavih s stališča (ne)kultiviranosti, ni (več) primeren, ker je bistveno prepočasen in premalo robusten, hkrati pa preveč elitistično zastavljen, da bi v okoliščinah 21. stoletja lahko služil slovenski jezikovni skupnosti, osvobojeni nevarnosti, da ji nadrejeni tujegovoreči večnacionalni zakonodajalec krati jezikovne pravice. Verjetno je v novih okoliščinah največja nevarnost, da bi govorci začeli opuščati slovenščino, ravno občutek izključenosti iz omenjenih globalnih trendov.

Nova radikalizacija, povezana s prikrito redefinicijo pojma knjižni jezik, ki jezikovno skupnost potiska v infantilizacijo s potrebo po nenehnem spraševanju avtoritarnega organa o lastnostih njihovega lastnega jezika, ne da bi ji ta iz samoohranitvenih razlogov sploh kdaj hotel ponuditi razumljiv koncept skupnega standarda, na katerega se lahko vsak trenutek opre brez pomoči interpreta, je dodatni razlog za prevpraševanje, kateri del svojega jezika želi ta skupnost standardizirati in kako. Za ta namen je treba na novo definirati pojem *standardnega jezika*, katerega značilnosti so predvsem:

- konceptualno je standardni jezik z upoštevanjem nove diskurzne realnosti opredeljen z znano definicijo kot »skupek vzporedno živečih, enakovrednih stilov« (Vidovič Muha 1972: 187), ki se zavestno omejuje na tiste dele jezikovne stvarnosti, ki jezikovni skupnosti služijo kot skupno mesto sporazumevanja v okviru formalnih sistemov, kot so izobraževalni sistem, sistem javne oz. državne uprave, vključno z umeščenostjo v širšo skupnost EU, z izhodiščem v funkcionalni dogovornosti, ki izhaja iz ustrezno interpretirane jezikovne rabe;
- ideološko se standardni jezik umešča v nadaljevanje standardnojezikovne kulture, ki je značilna za slovensko jezikovno skupnost, pri čemer za del jezika, ki ga razume in definira kot standardnega, vir dogovora išče v spremljanju jezikovne rabe s sprejemanjem konsenzualnih odločitev na podlagi maksime o prilagajanju standarda na delih, ki največ govorcem povzročajo največ težav, ne pa v kultiviranosti, ki strukturno omogoča

in celo spodbuja arbitrarnost odločitev vnaprej določene »kultivirajoče institucije« ali »kultivirajočega organa«;

- izvedbeno je pojem standardnega jezika usmerjen v zagotavljanje možnosti, da vsak govorec slovenščine vsak trenutek lahko sam interpretira vse okoliščine rabe kateregakoli izraza v odnosu do njegove standardnosti ali nestandardnosti, v smislu, kot je pojem standardnosti opredeljen v prvi alineji.

II Sodobni standardizacijski priročniki in slovar

Tehnološka izvedba sodobnega digitalnega slovarja

Bojan Klemenc, Marko Robnik-Šikonja, Luka Fürst, Ciril Bohak in Simon Krek

Abstract

An important component in a state-of-the-art digital Slovenian language dictionary is its technological framework, which is briefly presented in this paper. We view the dictionary as a multi-tier architecture, with a presentation tier, a middle application tier (a back-end application system with a component for semi-automatic data extraction) and a data tier. In its natural form, language data is multidimensional. In a printed dictionary, there is just the presentation tier, and many of the relations between the underlying data are difficult to access or may even be lost. In electronic dictionaries, however, there are no such restrictions. The data can be preserved in all its complexity and presented in various ways because there is a distinction between the data and its presentation. This separation is the key factor in integrating the various data sources (different corpora and external databases) into a unified database. Various users or programs can then query different parts of the database based on their interests and the presentation tier displays or returns the data on different levels of granularity. For each tier we present the structure and review some of the technological considerations which ensure that the extensibility, reliability and adaptability of the final solution are to a high standard.

Keywords: digital dictionary, multi-tier software architecture, presentation layer, relational database, data extraction

Ključne besede: digitalni slovar, večdelna programska arhitektura, predstavitveni nivo, relacijska baza podatkov, luščenje podatkov

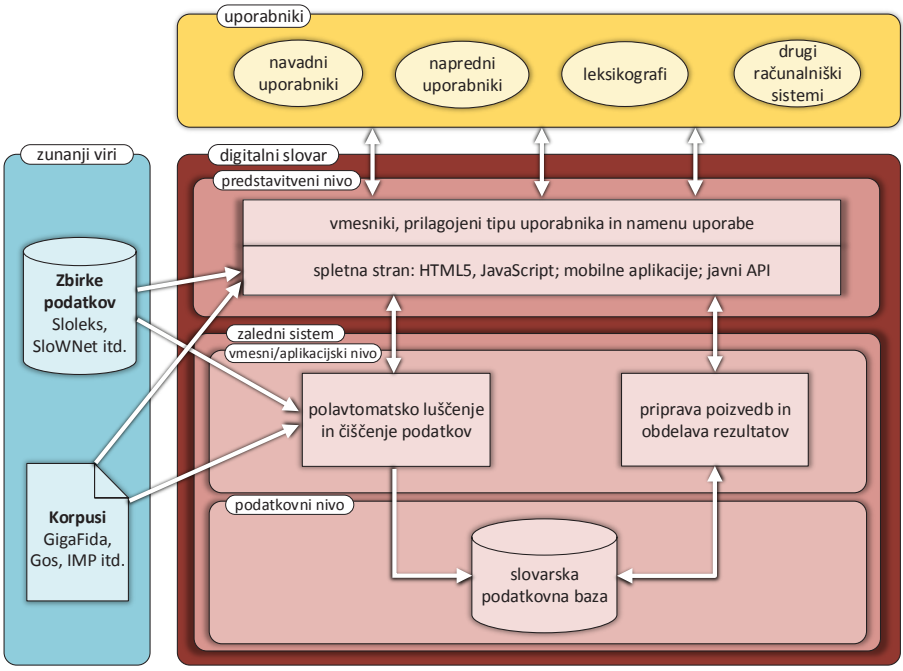
1 Uvod

Sodoben digitalni slovar slovenskega jezika bo imel poleg svoje vsebinske ravni tudi tehnološko, ki jo predstavljamo v tem prispevku. Najprej opišemo poglavitne komponente, ki so del tehnološke izvedbe takšnega slovarja, v nadaljevanju pa se prispevek dotakne tudi smernic za implementacijo takšnega slovarja. Pri tehnološki zasnovi slovarja je zelo pomembno, da so uporabljeni koncepti in tehnologije premišljeno izbrani z namenom trajnosti, razširljivosti, prilagodljivosti in zanesljivosti končne implementacije.

Prve generacije digitalnih oz. digitaliziranih slovarjev so bile z vidika podatkovnega modela zgolj preslikava obstoječih slovarjev v papirni obliki (prim. Urdang 1984, Boguraev in Briscoe 1989, Hajnšek-Holz 1993, Krek 2014): geselski članki so se s svojo hierarhično organizacijo in oznakami hranili v datotekah, npr. v datotekah XML (angl. *eXtensible Markup Language*) ali pri spletnih slovarjih neposredno v HTML (angl. *HyperText Markup Language*). V slednjem primeru sta logična struktura geselskega članka in njegova prikazna oblika (izgled) združeni, pri geselskih člankih v XML-u pa je podana struktura članka, medtem ko se prikazna oblika ustvari s pomočjo ustreznih transformacij po predlogi, kot je npr. CSS (angl. *Cascading Style Sheets*). V tem primeru že pridemo do osnovnega ločevanja med podatki in njihovim prikazom. Besedilo geselskega članka lahko vsebuje tudi reference na druge geselske članke oz. njihove sestavne dele. Poizvedbe, ki jih lahko izvajamo v takšnem slovarju, so tipično omejene na iskanje iztočnic, iskanje po določenih elementih (ki so tipično opredeljeni v XML-u) in splošno iskanje po besedilu geselskega članka. Prikaz rezultatov je v takšnih slovarjih vselej enak: geselski članek, morebiti z obarvanimi rezultati iskanja. Poizvedbi oz. iskanemu rezultatu bolj prikrojenih prikazov pa ne moremo dobiti, saj so podatki strukturirani za določeno število vnaprej definiranih prikazov in zato oblikovanje prikaza glede na trenutno povpraševanje ni mogoče. Takšna ureditev podatkov (geselskih člankov) je naravna, kadar imamo opravka z medijem, kot je papir, kjer morajo biti podatki že organizirani oz. shranjeni v svoji končni predstavitveni obliki. Digitalno zasnovani slovarji namreč nimajo te fizične omejitve, pri njihovem načrtovanju pa moramo preseči tako »nivo papirja« kot tudi statične podatkovne strukture. Podatke je potrebno hraniti v njihovi naravni večrazsežni obliki ter jih na podlagi želenih poizvedb ustrezno filtrirati, preurediti in prikazati.

Za izvedbo digitalnega slovarja je torej ključna ločitev predstavitve podatkov od podatkov samih. Podatke na ta način lahko hranimo v vsej njihovi kompleksnosti, predstavitev podatkov pa je glede na vso kompleksnost hranjenih podatkov mogoča z različnih zornih kotov in omogoča tako spremembo gledišča kot stopnjo podrobnosti predstavitve. Tehnološko gledano je pri tem glavna delitev na predstavitveni nivo in podatkovni nivo. Uporabniku podatkovni nivo ni viden in

do teh podatkov dostopa zgolj preko predstavitvenega nivoja. Predstavitveni nivo uporabniku prikazuje podatke in sprejema uporabnikove »poizvedbe« (klike, iskanja). Vez med obema nivojema predstavlja t. i. aplikacijski oz. vmesni nivo. Njegova naloga je, da poizvedbe predstavitvenega nivoja pretvori v obliko, s katero lahko od podatkovnega nivoja pridobi ustrezne podatke. Pridobljene podatke nato ustrezno prečisti in preoblikuje ter posreduje predstavitvenemu nivoju.



Slika 1: Shema arhitekturne delitve digitalnega slovarja na tri nivoje: predstavitveni, vmesni aplikacijski in podatkovni nivo. Uporabniki vidijo predstavitveni nivo (spletna stran, mobilne aplikacije), ki jim prikazuje ustrezno izbrane in obdelane podatke iz podatkovnega nivoja. Naloga vmesnega nivoja je, da povezuje podatkovni in predstavitveni nivo ter omogoča polnjenje slovarske baze iz zunanjih virov.

Tako dobimo trinivojsko arhitekturo (Slika 1), kjer je uporabniku viden zgornji predstavitveni nivo (angl. *presentation tier,* tudi *front end*), vmesni aplikacijski nivo in spodnji podatkovni nivo pa nista vidna, imenujemo ju tudi zaledni sistem (angl. back end). Plastovitost arhitekture omogoča, da so posamezni deli relativno neodvisni eden od drugega – višji nivoji preko vnaprej definiranih programskih vmesnikov dostopajo do nižjih nivojev. Posledično lahko zamenjamo posamezni nivo, ne da bi to negativno vplivalo na ostale nivoje. Ločevanje predstavitvenega nivoja od baze podatkov na podatkovnem nivoju omogoča integracijo slovarjev

in virov, ki so bili do sedaj ločeni, saj imamo enotno podatkovno bazo, iz katere črpajo različni »pogledi« predstavitvenega nivoja, da prikažejo podmnožico baze (npr. pisni jezik, govorjeni jezik, sodobni jezik, arhaični jezik, pokrajinske različice, kombinacije omejitev itd.).

Na predstavitvenem nivoju lahko prikazujemo različne vmesnike tudi glede na vrsto uporabnika – ena skupina uporabnikov slovarja lahko vidi/izbere drugačen vmesnik od druge, tako ima npr. srednješolec, ki uporablja slovar za pisanje eseja, popolnoma drugačen prikaz z drugimi in drugače hierarhiziranimi podatki kot jezikoslovec ali leksikograf. Vsi sicer dostopajo do iste baze podatkov, se pa razlikuje nivo podrobnosti vrnjenih podatkov oz. možnost vnašanja podatkov. Na primer leksikograf lahko podatke tudi spreminja, medtem ko jih drugi uporabniki ne morejo.

Posodabljanje slovarske baze se lahko izvaja tako s pomočjo ročnega dela leksikografa ali z množičenjem (angl. *crowdsourcing*) (prim. Kosem et al. 2013a; 2013b), sistem pa omogoča, da se podatki pripravljajo avtomatsko z luščenjem iz zunanjih virov (npr. korpusov). Luščenje ni zgolj enkratno opravilo, saj se jezik in posledično korpusi spreminjajo, tako da gre za ponavljajoč se proces. Vmesni aplikacijski nivo ima tako poleg naloge, da služi povezovanju predstavitvenega in podatkovnega nivoja, tudi nalogo, da se povezuje z zunanjimi viri in omogoča začetno avtomatsko luščenje podatkov.

Po tehnološki plati lahko slovar tako razdelimo na štiri poglavitne komponente, ki so na kratko predstavljene v nadaljevanju.

1. **Podatkovna baza** kot glavna komponenta podatkovnega nivoja je realizirana v obliki enotne relacijske podatkovne baze, ki je namenjena hrambi jezikovnih podatkov in iz korpusov izluščenih informacij.
2. **Zaledni aplikacijski sistem** oz. vmesni aplikacijski nivo je namenjen integraciji celotne rešitve in vsebuje programske vmesnike za dostop predstavitvenih modulov (spletna aplikacija, mobilne aplikacije) in programsko kodo za dostop do podatkovne baze.
3. **Komponenta za avtomatsko luščenje podatkov**, ki je v bistvu del vmesnega aplikacijskega nivoja in skrbi za polnjenje in ažurno obnavljanje baze podatkov iz zunanjih besedilnih korpusov in baz. Zaradi kompleksnosti jo bomo kot komponento obravnavali ločeno od preostalega aplikacijskega nivoja. Kot del leksikografskega procesa je avtomatsko luščenje podatkov predstavljeno tudi v Gantar et al. (2015).
4. **Predstavitveni nivo** v obliki spletnega portala s predstavitvijo vseh jezikovnih podatkov za različne tipe uporabnikov in mobilne aplikacije za različne mobilne platforme (npr. Android, Apple iOS in Windows Phone), ki omogočajo iskanje in brskanje po jezikovnih podatkih ter v opisanem nadzorovanem leksikografskem procesu (ibid.) tudi sodelovanje

pri popravljanju in dopolnjevanju jezikovnih podatkov. Uporabniki niso nujno samo ljudje, zato preko predstavitvenega nivoja izpostavimo tudi programski vmesnik, preko katerega lahko drugi računalniški sistemi dostopajo do slovarja.

Tehnološko izvedbo slovarja je smiselno v večji meri zasnovati na odprtokodnih rešitvah, ki so danes že dovolj zmogljive, da podpirajo tudi zahtevne operacije in veliko število uporabnikov. Pri izbiri tehnologij nam delitev na nivoje omogoča, da na vsakem nivoju izberemo najustreznejše tehnologije oz. jih po potrebi zamenjamo. Enako načelo velja tudi za posamezne komponente. Na primer: komponenta za avtomatsko luščenje podatkov je ločena od ostalih komponent na aplikacijskem nivoju; z njimi po potrebi komunicira preko programskih vmesnikov.

Komunikacija med posameznimi nivoji poteka po modelu odjemalec–strežnik. Odjemalec pošlje zahtevo strežniku, ta pa pošlje ustrezen odgovor. Odjemalci imajo v primeru slovarja lahko zaradi tega manjše procesorske in pomnilniške zahteve, saj se podatki v veliki večini hranijo na strežniku in se tam tudi obdelujejo, odjemalcu pa se pošljejo le podatki odgovora, ki jih odjemalec (predstavitvenega nivoja) potem ustrezno prikaže. Manjše procesorske in pomnilniške zahteve pomenijo manjšo porabo energije, kar omogoča uporabo slovarjev na manj zmogljivih mobilnih napravah pod pogojem, da imamo podatkovno povezavo do strežnika. Podatki v slovarski bazi se tako v procesu izdelave in tudi kasneje redno spreminjajo, zato je takšna arhitekturna rešitev primerna, saj imajo uporabniki vedno dostop do najbolj ažurne različice podatkovne baze. Vendar pa takšna arhitekturna rešitev ne pomeni, da morajo biti odjemalci in strežniki strogo nameščeni na različnih napravah, ampak so lahko tudi fizično na isti napravi. V tem primeru pride do replikacije (dela) baze, kar pomeni, da je treba poskrbeti, da so posamezne kopije baze ustrezno sinhronizirane (tipično z eno od kanoničnih kopij baze). Primer koristnosti takšne rešitve je, da lahko (tudi na mobilnih napravah) uporabljamo slovar brez povezave z internetom.

Večnivojska in modularna zgradba nam omogočata, da posamezne dele slovarja gradimo, evalviramo in testiramo vzporedno. Predpogoj za to pa je, da so povezave med posameznimi nivoji, npr. programski vmesnik, vnaprej dobro definirane.

2 Podatkovni model in baza podatkov

Enotna podatkovna baza in ločen predstavitveni nivo omogočata integracijo slovarjev in virov, ki so bili do sedaj ločeni. Da lahko zgradimo ustrezno enotno podatkovno bazo, je na eni strani treba definirati ustrezen podatkovni model, ki bo lahko hranil integrirane podatke iz različnih obstoječih in novonastalih baz. Poleg tega mora omogočati širši nabor poizvedb, da pokrije tiste, ki so se že izvajale na

obstoječih bazah, in omogoči nove na integriranih podatkih. Na drugi strani pa se z integracijo oz. enotno bazo poveča količina hranjenih podatkov, ki morajo biti še vedno hitro dostopni.

Tabela 1 za posamezne jezikovne podatke, ki bodo prikazani v uporabniškem vmesniku, prikazuje vir podatkov, predvideno umestitev v enotno podatkovno bazo in obstoječ trenuten format podatkov. Pri vključenosti v bazo je navedena umestitev neposredno v bazo ali referenca na zunanje vire, npr. korpuse. Več o povezanih slovarskih in korpusnih virih v Krek et al. (2013b).

Tabela 1: Prikazani podatki, njihovi viri, način vključenosti v podatkovno bazo in trenutni format. Oznake formatov so naslednje: TEI (Text Encoding Initiative), LMF (Lexical Markup Framework) in LBS (Leksikalna baza za slovenščino).

Prikazani podatki	Vir podatkov	Vključenost podatkov v bazo	Trenutni format
besedne zveze	izluščeni podatki	DA, kot leksikon	XML LBS
besedne zveze - konkordance	Gigafida (Korpus slovenskega jezika Gigafida)	NE, povezava na konkordančnik	-
besedne oblike	Sloleks (Slovenski oblikoslovni leksikon Sloleks)	DA, kot leksikon	XML LMF
sinonimi in prevodi v izbrane tuje jezike	sloWNet (Slovenski semantični leksikon sloWNet)	DA, kot leksikon	XML DEBDIC
zgodovina, besede	IMP (Korpus starejše slovenščine IMP)	DA, kot leksikon	XML TEI
zgodovina - konkordance	IMP (Korpus starejše slovenščine IMP)	NE, povezava na konkordančnik	-
govor, besede	Gos (Korpus govorjene slovenščine Gos)	DA, kot leksikon	(XML TEI - izvedba v projektu)
govor - konkordance	Gos (Korpus govorjene slovenščine Gos)	NE, povezava na konkordančnik	-
vizualizacija relacij	izluščeni podatki	DA	XML LBS
multimedija	WikiMedia, ...	DA, tudi kot zunanji viri	različni multimedijski formati
jezikovna statistika	Gigafida (Korpus slovenskega jezika Gigafida)	DA	-

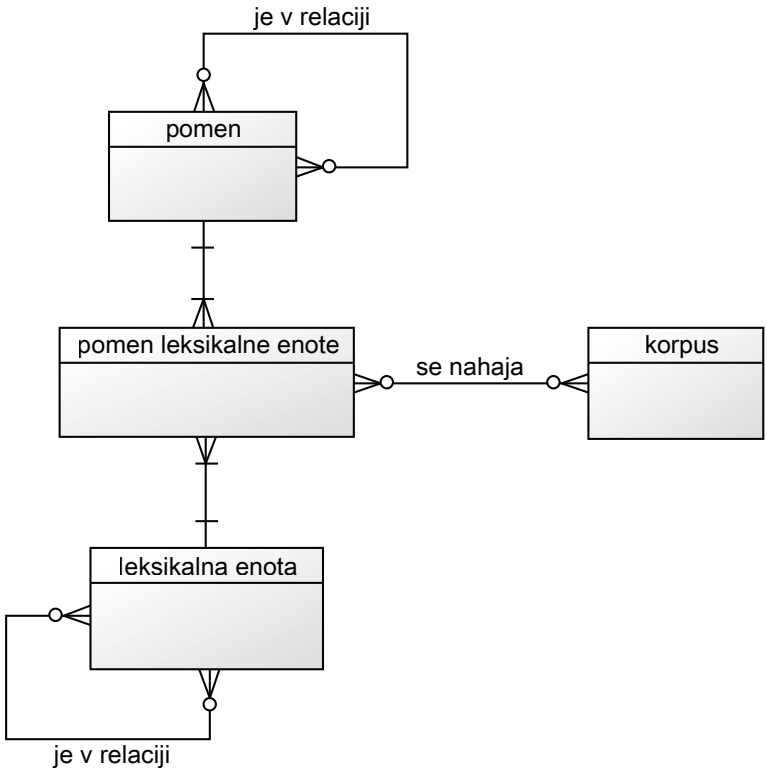
Če se osredotočimo na tekstovne oblike virov, so večinoma v zapisu XML ali v običajnih tekstovnih datotekah. XML poleg vsebinskih podatkov vsebuje tudi podatke o strukturi. Struktura podatkov v različnih virih ni enaka (tudi zaradi vsebine, ki jo pokrivajo), zato strukture XML v bazi ni smiselno ohranjati (obstajajo tudi izjeme, kjer je smiselno ohraniti manjše dele, npr. poudarki pri opisih). XML je po naravi hierarhična oblika hrambe podatkov, ki pa ni najbolj primerna za hranjenje podatkov, ki niso hierarhične narave (kot v primeru slovarjev). Vendar pa je oblika XML zaradi te hierarhičnosti precej primerna za serializacijo, poleg tega pa sama datoteka XML vsebuje podatke o strukturi podatkov. Zaradi teh dveh razlogov je primerna za izmenjavo podatkov (z zunanjimi viri in z zunanjimi aplikacijami, ki preko programskih vmesnikov dostopajo do slovarja).

Pri podatkih v slovarski bazi so pomembne medsebojne relacije med posameznimi zapisi. Za modeliranje teh relacij so primerne grafovske podatkovne baze in relacijske podatkovne baze. Glede na zmogljivosti (Vicknair et al. 2010) sta obe ti vrsti podatkovnih baz primerni za velike količine podatkov, ki se pojavljajo pri slovarjih. Obe vrsti imata definirane poizvedovalne jezike: pri grafovskih bazah npr. SPARQL (SPARQL Query Language for RDF) in več nestandardnih rešitev (Wood 2012, Haase et al. 2004), pri relacijskih podatkovnih bazah pa sta uveljavljena standarda SQL in SQL/PSM. Grafovske baze so precej fleksibilne, saj nimajo eksplicitno definirane strukture. Primerne so za podatke, ki imajo zelo variabilno strukturo. Relacijske podatkovne baze imajo eksplicitno definirano strukturo, zato je potrebno vnaprej dobro definirati podatkovni model. Hkrati tako tudi vnaprej načrtujemo, katere poizvedbe nad bazo so možne in katere ne. Vseeno lahko tudi relacijski model prilagajamo tako, da del strukture hranimo kot podatke (Newman 2007).

Multimedijski viri se hranijo kot referenca in pred vključitvijo podatkov v bazo ustrezno tekstovno označijo, tako da jih lahko lažje preiskujemo.

Slovarska baza je zaradi zrelosti tehnoloških rešitev načrtovana kot relacijska podatkovna baza. Poenostavljen konceptualni model jedra podatkovne baze je prikazan na Sliki 2. Leksikalna enota nosi en pomen ali več pomenov. Pomeni so lahko med seboj v različnih relacijah. Leksikalne enote so lahko leksemi, stalne zveze ali fraze, tudi deli besed in so lahko med seboj v različnih relacijah. Za leksikalne enote z določenim pomenom hranimo (agregirane) podatke o tem, v katerih virih smo jih našli.

Model je zasnovan dovolj splošno, da omogoča širitev obsega hranjenih podatkov na več jezikovnih zvrsti in jih obravnava enakovredno. Poleg tega odločitve pri gradnji podatkovnega modela določajo tudi stopnjo granularnosti podatkov (manjša



Slika 2: Poenostavljen konceptualni model jedra podatkovne baze v Martinovi notaciji, ki služi kot izhodišče za načrtovanje celotne baze.

stopnja granularnosti pomeni, da hranimo več agregiranih podatkov ali manj natančne podatke, posledično na določene poizvedbe ne bomo mogli odgovoriti). Granularnost je pomembna pri luščenju podatkov in polnjenju baze, ker določa, kaj vse je potrebno izluščiti in kakšno dodatno delo bo potrebno opraviti, npr. pri množičenju in končni leksikografski obdelavi. Na primer: če se pri luščenju podatkov za posamezne leksikalne enote ne beleži podatkov o časovnem razponu pojavljanja (recimo na podlagi pojavitev v korpusih v določenem obdobju), potem ne bo možno omejiti poizvedb na besedišče iz določenega obdobja.

Pri postavitvi podatkovne baze imamo na izbiro več sistemov za upravljanje s podatkovnimi bazami. Ker so relacijske podatkovne baze precej uveljavljene, obstaja več odprtokodnih rešitev, vendar vse nimajo potrebnih funkcionalnosti. Sistem za upravljanje s podatkovno bazo mora podpirati tudi t. i. rekurzivne poizvedbe in SQL/PSM (v bazi shranjene procedure), zato smo omejeni na sisteme, ki le-te podpirajo (npr. PostgreSQL[1]).

1 www.postgresql.org (dostop 12. 6. 2015).

3 ZALEDNI APLIKACIJSKI SISTEM

Zaledni aplikacijski sistem predstavlja vmesni sloj med podatkovnim nivojem in predstavitvenim nivojem. Avtomatsko luščenje podatkov je sicer del zalednega aplikacijskega sistema, vendar ga zaradi obsežnosti obravnavamo v posebnem razdelku. Naloga sistema je, da zahteve po podatkih, ki jih prejme od predstavitvenega nivoja, ustrezno (pre)oblikuje in pošlje podatkovni bazi oziroma zunanjim virom (korpusom, zunanjim bazam). Njihove odgovore ustrezno obdela, prečisti in posreduje predstavitvenemu nivoju. Tukaj je pomembno ločiti med podatki in dodatnimi omejitvami in pravili, ki jih definiramo nad podatki, vendar pa se lahko te omejitve in pravila s časom tudi spreminjajo. Na primer, kolokacije, ki se pojavljajo pri posamezni leksikalni enoti, lahko beležimo čez daljše obdobje. Če želimo privzeto izpisovati samo kolokacije, ki so se ob leksikalni enoti pojavljale v določenem časovnem obdobju, npr. od leta 2005 do 2015 (zadnjih 10 let), potem je definicija tega kolokacijskega obdobja pravilo – tekom časa se kolokacije spreminjajo. Naloga aplikacijskega nivoja je, da omogoča definiranje takšnih pravil in ustrezno formulira poizvedbe na podatkovni bazi glede na dana pravila in omejitve. To ne pomeni, da smo omejeni na obdobja, ki jih določajo omejitve; preko uporabniškega vmesnika lahko eksplicitno določimo obdobje (ali pa ustrezno označimo ostale kolokacije pomena).

Aplikacijski sistem svoje storitve ponuja v obliki programskega vmesnika. Prednost ločenih nivojev je, da se lahko programska koda aplikacijskega nivoja spreminja (dopolnjuje, popravlja, izboljšuje), medtem ko ostane programski vmesnik enak in lahko odjemalci predstavitvenega nivoja (spletna aplikacija, mobilne aplikacije) brez težav dostopajo do storitev. Poleg odjemalcev predstavitvenega nivoja je potrebno omogočiti dostop tudi drugim računalniškim sistemom, ki bi želeli dostopati do podatkov, in omogočiti povezljivost v smislu semantičnega spleta (povezani podatki, angl. *linked data*).

Ker komunikacija med predstavitvenim nivojem in aplikacijskim nivojem deluje po principu odjemalec-strežnik, je pomembna naloga aplikacijskega nivoja, da se podatki pripravijo tako, da odjemalci dobijo le tiste podatke, ki jih nujno potrebujejo, in ni nepotrebnih prenosov.

4 AVTOMATSKO LUŠČENJE PODATKOV

Podatki v slovarski bazi so izluščeni iz različnih zunanjih virov, ki jih prikazuje Tabela 1. Pri luščenju podatkov se srečamo z dvema poglavitnima problemoma: količina podatkov, ki predstavlja vir, iz katerega luščimo (npr. korpus Gigafida

vsebuje približno 1,2 milijardi besed), in kako zagotoviti kakovost izluščenih podatkov. Proces luščenja zaradi časovne dinamičnosti jezika ni zaključen z objavo slovarja, ampak je trajen proces. Zaradi zgornjih zahtev se luščenje izvaja v prvem fazi avtomatsko, rezultate te faze ustrezno ovrednotimo in tiste z veliko stopnjo zanesljivosti vpišemo v bazo, rezultate z manjšo zanesljivostjo pa pošljemo v fazo čiščenja in ročnega obdelovanja.

Pri avtomatski fazi izhajamo iz metod luščenja podatkov, ki so bile razvite za potrebe sestavljanja Leksikalne baze za slovenščino v okviru projekta Sporazumevanje v slovenskem jeziku (Gantar 2009; Gantar in Krek 2011) in jih nadgradimo z novimi spoznanji in tehnološko izboljšanimi orodji. Za celotno besedišče, ki bo vizualizirano, se lahko strojno pridobi naslednje podatke: iztočnico v osnovni obliki, besedno vrsto, podatek o pogostosti v korpusu, slovnične relacije, ki se v bazi prepišejo v vzorce, ter pripadajoče kolokacije in njihovi zgledi. Za postopek avtomatizacije je že bila izdelana t. i. slovnica besednih skic, ki deluje v orodju Sketch Engine.[2] S pomočjo prilagojene programske skripte, ki vsebuje opise vseh relevantnih slovničnih relacij za luščenje kolokacij, t. i. konfiguracije GDEX (okrajšava za angl. *Good Dictionary Examples*), ki opredeli lastnosti dobrih zgledov, lahko iz korpusov avtomatsko pridobimo čim boljše kandidate za primere uporabe posameznih iztočnic v realnem besedilnem okolju (Kosem et al. 2011).

V drugi fazi se podatki pred vključitvijo v slovarsko bazo ročno pregledajo. Delo se opravlja s pomočjo množičenja, kjer uporabniki označujejo, ali so v rezultatih anomalije oz. napake. Na koncu podatke preuredi in potrdi leksikograf. Potrjene napake, ki so posledica avtomatskega luščenja, se označijo in vračajo kot informacija nazaj v sistem luščenja, ki se iz njih uči z uporabo tehnik strojnega učenja in tako izboljšuje svoje delovanje.

Avtomatsko luščenje podatkov spada v zaledni sistem. Delno in končno obdelane podatke zapisujemo v slovarsko podatkovno bazo. Vsi podatki, ki niso dokončno obdelani, so v bazi ustrezno označeni, kar pomeni, da jih lahko na predstavitvenem nivoju bodisi prikažemo bodisi ne prikažemo. Na primer, leksikograf in splošni uporabnik dostopata do iste baze, vendar bo leksikograf poleg drugačnega uporabniškega vmesnika videl tudi podatke, ki niso dokončno obdelani, in jih lahko ustrezno obdeloval. Tudi uporabniki, ki sodelujejo v množičenju, imajo svoj pogled na podatke. Za množičenje se lahko uporabijo obstoječe platforme, kot je npr. PyBossa[3], ki omogočajo preprostejšo izdelavo aplikacij za množičenje (prim. Fišer et al. 2015).

2 http://www.sketchengine.co.uk/ (dostop 12. 6. 2015).
3 http://pybossa.com/ (dostop 12. 6. 2015).

5 PREDSTAVITVENI NIVO: SPLETNI PORTAL IN MOBILNE APLIKACIJE

Predstavitveni nivo mora zaznamovati uporabniška izkušnja in s tem posledično ustrezen uporabniški vmesnik aplikacij, kar ima velik vpliv na njihovo uspešnost. Zelo pomembna je tudi celostna grafična podobna aplikacij. Namen predstavitvenega nivoja je tudi v kar najbolj podobni obliki prikazovati informacije na spletnih straneh in priljubljenih mobilnih platformah.

Za razvoj mobilnih aplikacij je smiselno uporabiti t. i. hibridni pristop, ki predstavlja najboljši način za prenosljivost aplikacij med različnimi mobilnimi platformami pri čim višji ponovni uporabljivosti posameznih razvitih delov. Pri tem je trenutno za razvoj osnovnih funkcionalnosti smiselno uporabiti tehnologiji HTML5 in Javascript. Na tak način razvito jedro aplikacije lahko nato umestimo v aplikacijsko ogrodje posamezne podprte platforme. Takšen razvoj podpirajo številna odprtokodna orodja, npr. PhoneGap,[4] ki temelji na platformi Apache Cordova.[5] To olajša in pospeši razvoj aplikacij za vse podprte platforme, zagotavlja pa tudi enoten predstavitveni nivo na vseh platformah, kot tudi poenostavljeno posodabljanje aplikacij. Osnova tako razvite mobilne aplikacije je lahko osnova pri razvoju spletnega portala.

Z namenom prepoznavnosti in enotne uporabniške izkušnje pri uporabi aplikacij je smiselno zasnovati celostno grafično podobo uporabniškega vmesnika. Pomembno je, da zasnova upošteva standard WCAG 2.0 (Web Content Accessibility Guidelines 2.0), s čimer je omogočena tudi raba uporabnikom s posebnimi potrebami.

6 ZAKLJUČEK

Pri tehnološki izvedbi sodobnega digitalnega slovarja slovenskega jezika je ključna ločitev predstavitve podatkov od podatkov samih. Podatke na ta način lahko hranimo v vsej njihovi kompleksnosti, predstavitev podatkov pa je mogoča z različnih zornih kotov in omogoča tako spremembo gledišča kot stopnjo podrobnosti predstavitve. Arhitekturno je tehnološka izvedba zasnovana trinivojsko, kjer imamo predstavitveni nivo, vmesni aplikacijski nivo in podatkovni nivo. Naloga predstavitvenega nivoja je, da uporabniku prikaže podatke, ki so shranjeni na podatkovnem nivoju. Med obema nivojema je vmesni aplikacijski nivo, ki pretvori uporabnikove poizvedbe s predstavitvenega nivoja v obliko, ki je primerna za poizvedovanje direktno na podatkovnem nivoju (podatkovni bazi). Na drugi strani pa vmesna aplikacijska plast preoblikuje podatke v obliko, ki jo potrebuje predstavitveni nivo. Še ena funkcionalnost vmesnega aplikacijskega nivoja je

4 http://phonegap.com/ (dostop 12. 6. 2015).
5 https://cordova.apache.org/ (dostop 12. 6. 2015).

avtomatizirano luščenje podatkov iz korpusov in zunanjih zbirk podatkov. Ker se jezik razvija, je avtomatizirano luščenje podatkov stalen proces, pri katerem sodelujejo tudi leksikografi, ki dostopajo do podatkov preko ustreznih prikazov predstavitvenega nivoja. Ločitev med podatki in predstavitvijo je ključen dejavnik pri integraciji različnih virov (korpusov in zunanjih zbirk podatkov) v enotno podatkovno bazo. Različni uporabniki ali pa tudi zunanji računalniški sistemi nato s pomočjo poizvedb, posredovanih preko predstavitvenega nivoja, pridobijo želene podatke iz podatkovne baze, ki jih potem prikaže predstavitveni nivo.

Prednost nivojske arhitekture je neodvisnost posameznih nivojev, dokler je programski vmesnik, preko katerega višji nivoji dostopajo do nižjih, ustrezno definiran. Na vsakem nivoju lahko zato izberemo najustreznejše tehnologije za izvedbo in sprememba na enem od nivojev nima večjega vpliva na ostale nivoje, dokler se programski vmesnik ne spreminja. Tehnološko zasnovo slovarja lahko z upoštevanjem nivojske arhitekture razdelimo na 4 komponente: podatkovno bazo (podatkovni nivo), zaledni aplikacijski sistem s komponento za delno avtomatizirano luščenje podatkov (oboje spada v vmesni aplikacijski nivo, vendar obravnavamo komponento za luščenje podatkov ločeno zaradi njenega obsega in pomena za celoten sistem) in predstavitveni del, kjer imamo spletni portal in mobilne aplikacije (predstavitveni nivo).

Opisana tehnološka zasnova slovarja zagotavlja, da na njej sloneča rešitev služi kot osrednji interaktivni spletni jezikovni portal z opisom vseh ravnin besedišča slovenskega jezika. Komponente omogočajo trajnostni razvoj portala, namenjenega tako uporabnikom spleta kot mobilnih naprav in bi bile pod prosto dostopno licenco na voljo za nadaljnji razvoj.

Oblikoslovne informacije v sodobnih slovarskih priročnikih

Kaja Dobrovoljc

Abstract
Although morphology in lexicography is generally considered to be a solved problem which mostly deals with user-oriented evaluations of its comprehensibility, online dictionaries bring new possibilities for both dictionary users and dictionary makers alike. In the context of planning a dictionary of contemporary Slovenian, this paper explores the language users' need for morphological information, and the different aspects of its inclusion in a born-digital online dictionary. Preliminary analysis of inflection dictionary log files confirms that there is a great need for the inclusion of inflectional information, and that users tend to search for both regular and irregular inflectional paradigms. However, this need is not sufficiently met within the recently issued edition of reference, the Dictionary of Literary Slovenian, as decoding inflectional and other morphological information requires substantial cognitive effort and metalinguistic knowledge that cannot be expected from most users. Given that Slovenian is a morphologically rich language with extensive inflectional information, we take into account the idea of a separate machine-readable morphological database intended for use in language guides and various NLP applications. This database brings many advantages for dictionary users, such as the display of full inflectional, pronunciation and derivational paradigms, normative information, hyperlinking, improved searching, corpus linking, speech synthesis and voice search recognition. At the same time, it demands careful consideration of the content-related, visual and technical issues that arise when interlinking two distinct databases, in particularly morphology-dependent polysemy and variant spelling synonymy.

Keywords: morphology, inflection, morphological lexicon, dictionary database

Ključne besede: oblikoslovje, pregibanje, oblikoslovni leksikon, slovarska baza

1 UVOD

Slovarji poleg pomenskih lastnosti besedišča običajno prinašajo tudi informacije o njihovih izraznih lastnostih, kot so podatki o izgovoru, pregibanju, zapisovanju in drugih oblikoslovnih informacijah, ki za razliko od recepcijsko usmerjenih semantičnih informacij slovarskim uporabnikom lajšajo rabo leksikalnih enot v procesu besedilne produkcije. To velja tudi za slovenski prostor, saj se vse od prve izdaje Slovarja slovenskega knjižnega jezika tako v splošne kot specializirane, terminološke, zgodovinske, narečne in druge enojezične slovarje kot nepogrešljiv del slovarskega opisa praviloma vključujejo tudi informacije o izgovoru, pregibanju in oblikoslovnih lastnostih obravnavanih iztočnic.

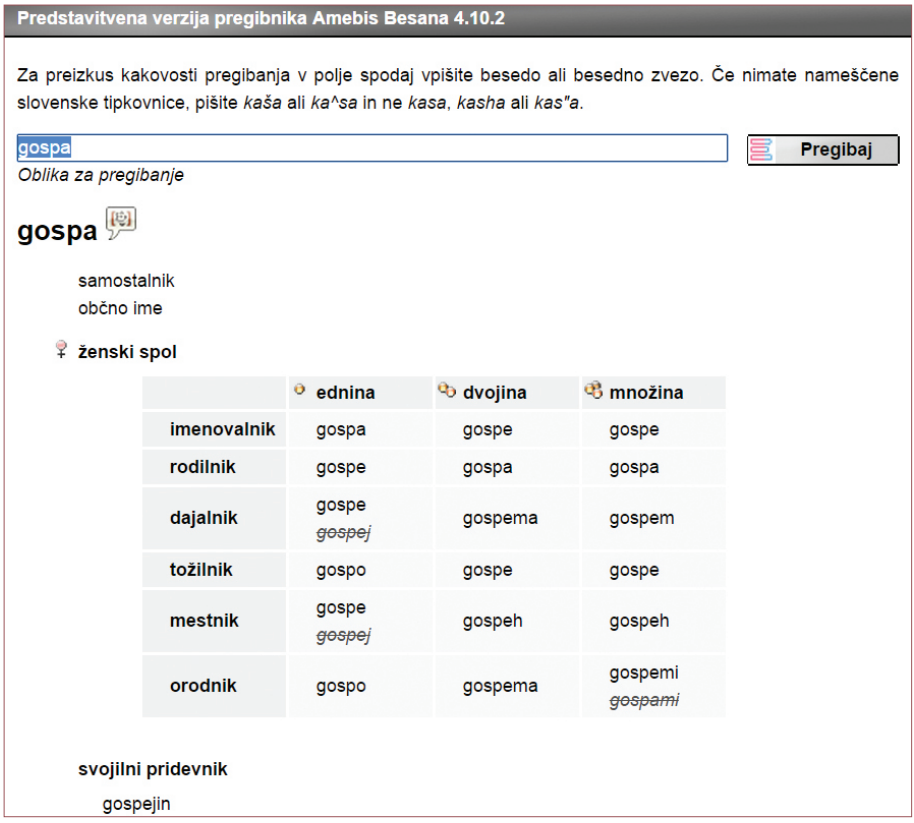
Kljub inherentnosti oblikoslovnih informacij v leksikalnih opisih jezika pa je bilo tako v slovenski kot širši leksikografski stroki vprašanjem, povezanim z opisom izraznih lastnosti besedišča v slovarjih doslej namenjeno razmeroma malo pozornosti. Medtem ko je pri morfološko manj kompleksnih jezikih v kontekstu tiskano zasnovanih slovarjev prevladovalo predvsem vprašanje, ali uporabniki z izjemo sistemsko neregularnih ali nepredvidljivih posebnosti sploh potrebujejo informacijo o pregibanju besed ter v kolikšni meri je ta predvidljiva pri tujejezičnih govorcih (Jackson 2002: 105–107; Honselaar 2003: 355–356; Caluwe in Santen 2003: 73–77), so se tudi razprave morfološko kompleksnejših jezikov osredotočale predvsem na mikrostrukturni vidik najučinkovitejšega podajanja pregibnih informacij, kot so način in razumljivost krajšanja pregibnih oblik ali kodnega sklicevanja na paradigmatske vzorce (Vikør 2009: 140; Kola 2012). Pri tem je v slovenskem prostoru bolj kot o načinu posredovanja oblikoslovnih informacij prevladovala predvsem razprava o njihovi ustreznosti z vidika knjižnojezikovne norme (prim. Toporišič 1971a; 1971b; Rigler 1971; 1972).

Ob dejstvu, da spletni medij tudi z vidika oblikoslovnih slovarskih informacij prinaša številne nove možnosti njihovega beleženja in posredovanja, v kontekstu načrtovanja digitalno zasnovanega slovarja sodobne slovenščine v pričujočem prispevku v prvem delu z analizo iskanj po pregibniku Besana empirično utemeljimo potrebo po vključevanju oblikoslovnih informacij[1] v bodoče slovarske opise slovenskega jezika (2) in analiziramo njeno zadovoljitev v obstoječem referenčnem slovarskem priročniku za slovenščino (3). Na podlagi strokovno usklajenega predloga beleženja oblikoslovnih informacij v obliki ločene, s slovarsko bazo povezane podatkovne zbirke, v osrednjem delu nato nakažemo nekaj prednosti, ki bi jih tovrstna rešitev lahko ponudila slovarskim uporabnikom (4.1), obenem pa izpostavimo potrebo po leksikografskem zavedanju ločnice med podatki v oblikoslovni bazi in njihovem prikazovanjem v slovarju (4.2) ter med iztočnicami v oblikoslovni in slovarski bazi (4.3).

1 V prispevku se osredotočimo predvsem na informacije o pregibanju.

2 POTREBE UPORABNIKOV

Kot izhodišče razmislekov o uporabniških potrebah po vključevanju oblikoslovnih informacij v spletne jezikovne priročnike v nadaljevanju predstavimo pilotno analizo dnevniških datotek poizvedb po spletnem pregibniku podjetja Amebis,[2] ki je bil razvit kot eden izmed modulov slovničnega pregledovalnika Besana (Holozan 2012). Predstavitvena različica pregibnika je zasnovana kot spletni vmesnik, ki uporabnikom za iskanje po besedi ali besedni zvezi vrne podatek o njenih slovničnih lastnostih, morebitnih standardnih in nestandardnih pregibnih oblikah ter besedotvorno povezanih informacijah (Slika 1). Pregibnik temelji na leksikalni podatkovni zbirki ASES (Arhar in Holozan 2009), ki se sproti nadgrajuje in trenutno vsebuje več kot 244.000 leksikalnih enot.



Slika 1: Primer iskanja v pregibniku Amebis Besana.

V analizo so bile zajete dnevniške datoteke iskalnih poizvedb v obdobju šestih let (od januarja 2009 do januarja 2015), v katerih so bile beležene informacije

2 http://besana.amebis.si/pregibanje/ (dostop 30. 6. 2015).

o vnesenem iskalnem pogoju oz. nizu in številu takih iskalnih poizvedb. Kot prikazujejo podatki v Tabeli 1, je bilo v navedenem obdobju v spletnem servisu izvedenih 2.350.778 poizvedb za 787.751 različnih iskalnih pogojev. To pomeni, da je sistem dnevno v povprečju zabeležil več kot 1.000 iskalnih poizvedb,[3] kar ob dejstvu, da je pregibnik Besana le eden izmed spletnih oblikoslovnih priročnikov za slovenščino,[4] potrjuje, da uporabniška potreba po tovrstnem tipu jezikovnih informacij ni nezanemarljiva.

Tabela 1: Število iskalnih poizvedb po pregibniku Amebis Besana v obdobju 2009–2015.

Oblika iskalne poizvedbe	Št. vseh poizvedb	Št. različnih poizvedb
Beseda	2.250.705	723.608
Besedna zveza	100.073	64.143
SKUPAJ	**2.350.778**	**787.751**

Da bi odgovorili na vprašanje, po katerih besedah ali besednih zvezah uporabniki najpogosteje poizvedujejo in na kakšen način, smo v nadaljnjo kvalitativno analizo zajeli seznam tistih iskalnih nizov, po katerih so uporabniki v danem obdobju poizvedovali vsaj 300-krat. Čeprav je teh skupno le 571, predstavljajo več kot četrtino vseh iskalnih poizvedb (619.117 iskanj), kar kaže na to, da je pregibanje določenih leksikalnih enot za govorce slovenščine izrazito bolj problematično.

Podatki v Tabeli 2 prikazujejo, da med njimi prevladujejo predvsem občni samostalniki, kot so *hiše*[5] (26.115 poizvedb), *otrok* (22.164), *dan* (15.488), *hči* (14.046), *mati* (10.824), *gospa* (10.756), *človek* (6.941), *tla* (6.006), *otroci* (4.838), *vodja* (4.782), *pljuča* (4.501), *vrata* (4.408), *drva* (4.199), *oko* (4.034), *hiša* (3.957), *dno* (3.333), *pes* (3.296), *breskev* (3.032), *okno* (2.991), *leto* (2.967). Sledijo jim glagoli, npr. *zvedeti* (4.426), *dati* (3.401), *biti* (3.394), *iti* (3.201), *jesti* (2.259), *imeti* (1.736), *vedeti* (1.474), *moči* (1.100), *delati* (1.090), *poslati* (1.076); zaimki, npr. *on* (4.325), *nič* (3.518), *jaz* (3.334), *ta* (3.292), *ona* (3.059), *kaj* (2.473), *kar* (2.205), *kateri* (2.079), *ti* (1.901), *moj* (1.892); in lastna imena, npr. *Oselica* (20.731), *Miha* (5.271), *Luka* (5.120), *Marko* (3.115), *Jaka* (2.310), *Žiga* (2.144), *Mitja* (2.046), *Grosuplje* (1.985), *Sašo* (1.722), *Klemen* (1.598). Bistveno redkeje uporabniki poizvedujejo po pridevnikih, npr. *lep* (1.183), *nov*

3 Za primerjavo navedimo, da je bilo na spletnem portalu Slovenskega pravopisa v zadnjih petih letih v povprečju dnevno zabeleženih manj kot 400 poizvedb.

4 Med prosto dostopnimi spletnimi priročniki podoben način poizvedovanja po celotnih pregibnih paradigmah omogoča še leksikon Sloleks (dostopen na portalih http://www.slovenscina.eu/sloleks in http://www.termania.net, dostop 30. 6. 2015), podatki o pregibanju pa so v okrajšani obliki vsebovani tudi v večini slovarskih priročnikov Inštituta za slovenski jezik Frana Ramovša (dostopnih na portalih http://www.fran.si, http://bos.zrc-sazu.si/ in http://www.termania.net , dostop 30. 6. 2015).

5 Na seznamih najpogostejših iskalnih poizvedb naštevamo tudi tiste, pri katerih lahko na podlagi pojasnila avtorjev visoko pogostost pojavljanja pripišemo dejstvu, da so na portalu sugerirane kot vzorčne iskalne poizvedbe ali so bile vnesene pri preverjanju delovanja sistema, npr. *hiše*, *hiša* ali *Oselica*.

(690), *dober* (686), številkih, npr. *dva* (2.530), *tri* (1.500), *dve* (921), in prislovih, npr. *lahko* (685), *dobro* (593), *rad* (562), kar kaže na to, da je zaradi bolj sistemskega pregibanja teh besednih vrst z njimi povezanih manj uporabniških jezikovnih zadreg. Besednih zvez med analiziranimi iskalnimi pogoji ni, saj tudi najpogostejša (*dve leti*) ne dosega izbranega frekvenčnega praga (241 poizvedb).

Tabela 2: Najpogostejše iskalne poizvedbe glede na besedno vrsto.

	različnih poizvedb	vseh poizvedb
občno	336	398.658
glagol	71	53.633
zaimek	64	72.247
lastno	63	66.681
pridevnik	14	6.878
številnik	10	9.003
prislov	7	3.344
drugo	6	8.673
SKUPAJ	**571**	**619.117**

Pričakovano se med najpogostejšimi poizvedbami pojavljajo najbolj znane sklanjatvene in spregatvene posebnosti, kakršne so bile ugotovljene tudi pri analizi z jezikom povezanih vprašanj na jezikovnih forumih (H. Dobrovoljc in Krek 2011; Bizjak Končar et al. 2011) ali najpogostejših jezikovnih težav učencev (Kosem et al. 2012), po drugi strani pa analiza iskalnih poizvedb razkriva presenetljivo visok delež poizvedovanj po pregibanju navidezno neproblematičnih besed, ki v obstoječih priročnikih zaradi svojega sistemskega pregibanja doslej niso bile deležne posebne obravnave, npr. *avto* (2.034), *mama* (1.578), *miza* (1.565), *stol* (1.319), *fant* (1.070), *ura* (922), *knjiga* (918); *delati* (1.090), *videti* (776), *hoditi* (744), *govoriti* (612), *dobiti* (494); *lep* (1.183), *nov* (690), *prvi* (447), *star* (368), *zanimiv* (309) itd.

Čeprav v okviru pričujoče raziskave niso bili na voljo drugi potencialno relevantni metapodatki o iskalnih poizvedbah, kot so čas ogleda strani, število poizvedb ali lokacija uporabnika, ki bi lahko povedali več o profilu uporabnikov in uspešnosti njihove iskalne poizvedbe (prim. Müller-Spitzer et al. 2015), rezultati pilotne analize dnevniških datotek po pregibniku Besana potrjujejo, da med uporabniki obstaja realna potreba po vključevanju oblikoslovnih informacij v bodoče slovarske in druge priročnike za slovenščino, ter obenem kažejo, da ta ni omejena zgolj na razmeroma zaprt nabor že znanih oblikoslovnih posebnosti, temveč tudi na besedišče, ki se pregiba po sistemsko predvidljivejših vzorcih.

3 OBLIKOSLOVNE INFORMACIJE V SSKJ2

Druga, dopolnjena in deloma prenovljena izdaja Slovarja slovenskega knjižnega jezika (SSKJ2) v svojem uvodu izpostavi, da poleg pomenskih prinaša tudi informacije o izraznih lastnostih slovenskega besedišča, saj »*o vsaki besedi pove, kako se piše in izgovarja, kakšen dinamični in tonemski naglas ima, kako se pregiba, katere pomene ima in kakšni so odnosi med pomeni*« (Gliha Komac et al. 2014: 25). Pri tem slovar tako v tiskani kot spletni različici nadaljuje s tradicijo prve izdaje Slovarja slovenskega knjižnega jezika (SSKJ), v katerem je bila informacija o pregibanju slovarske iztočnice podana s kombinacijo podatkov v glavi oz. zaglavju ter njihove interpretacije s pomočjo pojasnil v uvodu. Uporabniki slovarja se morajo tako pred pridobivanjem informacije o pregibanju določene besede v uvodu najprej seznaniti s samim načinom podajanja teh informacij, obenem pa morajo za uspešno rešitev svoje jezikovne zadrege poznati tudi v uvodu pojasnjena načela njihovega tolmačenja. Ta lahko v splošnem opišemo kot štiristopenjski proces, ki ga sestavljajo (i) faza lematizacije (Uvod v SSKJ2: §27–§29), (ii) faza določitve besedne vrste (§30), (iii) faza razvezave krajšave druge oz. tretje osnovne oblike (§160–§165) in (iv) faza uvrstitve v ustrezno oblikoslovno-naglasno shemo (§180–§196).

Čeprav se zdi faza določanja relevantne iztočnice razmeroma neproblematična, analiza dnevniških datotek v drugem razdelku kaže, da uporabniki pogosto iščejo tudi po nekanoničnih oblikah, zato uspešnost pridobivanja pregibnih informacij v slovarju ne bi smela biti pogojena s poznavanjem načel določanja osnovne oblike iztočnic. Poleg poizvedb po oblikoslovno težavnih pregibnih oblikah, kot so npr. *gospe, hčer, dni/dnevi, njih, matere, brki, starš, sabo/seboj*, so med njimi namreč tudi take, pri katerih lahko predpostavljamo, da je uporabnik svojo poizvedbo skušal vnesti v abstrahirani osnovni obliki, a je pri tem izbral obliko v neustreznem zapisu, npr. *imati* ali *pluča*, ali slovnični kategoriji, npr. *psi, smuči, dve, ona, vsi, midve, onidve* ipd. Z vidika SSKJ2 je to problematično zlasti pri rabi spletne različice,[6] saj iskanje po variantnem zapisu iztočnice ali njeni pregibni obliki vrne rezultat zgolj, če se ta pojavi kot del geselskega sestavka (npr. brez zadetkov za poizvedbo *pluča*), pri čemer širitev na iskanje po celotnem sestavku in ne zgolj po iztočnici obenem vrne vse relevantne zadetke (npr. 78 gesel za poizvedbo *psi*).

Prav tako je lahko potencialno zahtevno tudi določanje besedne vrste in drugih slovničnih kategorij, ki jih uporabnik potrebuje pri izbiri ustreznih navodil za razvezavo krajšav, saj se te nahajajo v različnih delih geselskega članka, npr. s kvalifikatorsko oznako besedne vrste ali ene izmed lastnosti v glavi (npr. ***finále*** *-a m (â)* ali ***zanimív*** *-a -o prid.*), v razlagi (***sêstrin*** *-a -o (ē) svojilni pridevnik od sestra*

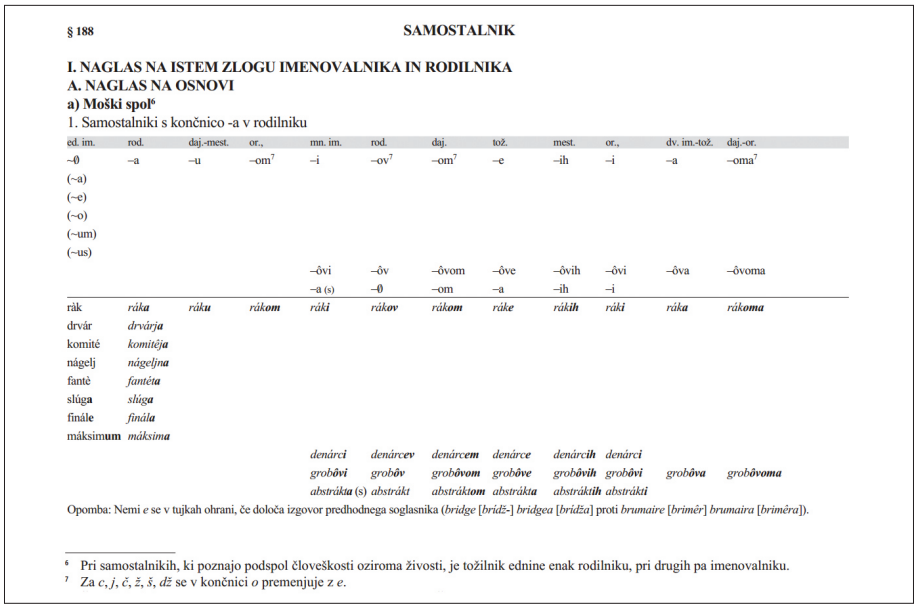
6 http://www.sskj2.si (dostop 30. 6. 2015).

ali ***bíl***[2] *-à -ó in -ò opisni deležnik od biti sem (î ä ọ̑)*), v kvalifikatorskem pojasnilu (***ánglo-*** *prvi del zvez (â)* ali ***si***[2] *članica*), v drugi iztočnici (***aloa*** *gl. aloja*) ali pa tega podatka v slovarju niti ni (***kamen...*** *prim. kamn...* in ***kamn...*** *prim. kamen...)*. Ko je uporabniku znana osnovna oblika iztočnice in podatek o njeni besedni vrsti, lahko nato v ustreznem paragrafu slovarskega uvoda poišče navodilo za razvezavo njene okrajšane druge oz. (pri pridevnikih) tretje osnovne oblike, ki pa zahteva zelo visoko stopnjo metajezikovnega znanja, kakršne ni mogoče pričakovati od laičnih uporabnikov ali govorcev, ki se jezika šele učijo, npr.:

> Pri samostalniku in pridevniških besedah je krajšanje takole: a) Kadar se prva oblika končuje na soglasnik, se druga oblika dela tako, da se prvi obliki doda zapisani del druge oblike, obstoječ iz samoglasnika ali j, n + samoglasnika /.../. Enako se dela druga oblika, kadar je zapisani del druge oblike -ih /.../. Če je zaradi sprememb v končniškem delu besede zapisan daljši del druge oblike ali pred končnico tudi kak soglasnik (razen j, n), je iz zapisa razvidno, na kateri del besede se nanaša /.../. b) Kadar se prva oblika končuje na samoglasnik, se druga oblika dela tako, da se prvi obliki doda zapisani del druge oblike, obstoječ iz j, t, n + samoglasnika /.../, ali pa se zadnji samoglasnik prve oblike izpusti, če se zapisani del druge oblike začenja s samoglasnikom /.../. Prav tako se dela druga oblika samostalnika s končnico -ega, ki se sicer sklanja po pridevniški sklanjatvi /.../. Če se zapisani del druge oblike začenja s kakim soglasnikom (razen j, t, n), je iz zapisa razvidno, na kateri del besede se nanaša /.../ (Uvod: § 161)

Vprašljiva je tudi razumljivost informacije o pregibanju kazalčnih iztočnic, kadar te nimajo enake stranske oblike kot njihova normativno ustreznejša dvojnica (npr. ***croquis*** *gl. kroki*, ki se ne pregiba enako kot ***krokí*** *-ja m (î)*) ter iztočnic, ki krajšave druge oblike v geselskem sestavku nimajo navedene (npr. ***múlda*** *ž (ú) jarek za odtok tekočine s ceste, tlakovanih površin* ali ***rímokatoličánka*** *ž (ī-â) pripadnica rimskokatoliške vere*).

V zadnji fazi razvezave uporabnik na podlagi kombinacije osnovne in razvezane druge oz. tretje osnovne oblike iztočnico uvrsti v ustrezno oblikoslovno-naglasno shemo (Slika 2), ki pri določanju tipa prav tako zahteva dobro poznavanje jezikoslovne terminologije (npr. *naglas na osnovi/končnici*, *naglas na različnih zlogih osnove*, *kratek/dolg naglas* itd.), pri uvrstitvi in razvezavi glede na podtip pa tudi upoštevanje posebnosti in modifikacij v opombah in pomen posebnih grafičnih znakov, kot so znak ~, ki označuje tvorbo iz imenovalniške in nedoločniške osnove oziroma dela osnove, znak –, ki označuje tvorbo iz rodilniške in sedanjiške osnove oziroma dela osnove (pri pridevniku pa se znaka nanašata na imenovalnik moškega oz. ženskega spola) ter znak ’, ki označuje naglasno mesto.



§ 188 SAMOSTALNIK

I. NAGLAS NA ISTEM ZLOGU IMENOVALNIKA IN RODILNIKA
A. NAGLAS NA OSNOVI
a) Moški spol[6]
1. Samostalniki s končnico -a v rodilniku

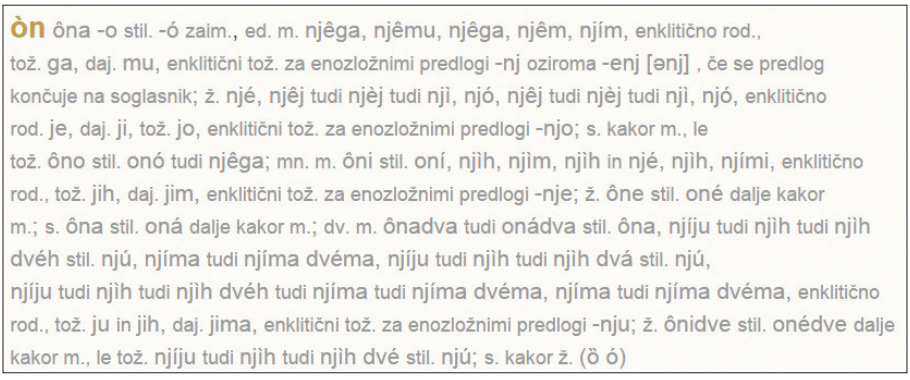
ed. im.	rod.	daj.-mest.	or.,	mn. im.	rod.	daj.	tož.	mest.	or.,	dv. im.-tož.	daj.-or.
–Ø	–a	–u	–om[7]	–i	–ov[7]	–om[7]	–e	–ih	–i	–a	–oma[7]
(~a)											
(~e)											
(~o)											
(~um)											
(~us)											
				–ôvi	–ôv	–ôvom	–ôve	–ôvih	–ôvi	–ôva	–ôvoma
				–a (s)	–Ø	–om	–a	–ih	–i		
rák	*rák**a***	*rák**u***	*rák**om***	*rák**i***	*rák**ov***	*rák**om***	*rák**e***	*rák**ih***	*rák**i***	*rák**a***	*rák**oma***
drvár	*drvárj**a***										
komité	*komitéj**a***										
nágelj	*nágelj**a***										
fantè	*fantét**a***										
slúg**a**	*slúg**a***										
finále	*finál**a***										
máksim**um**	*máksim**a***										
				*denárc**i***	*denárc**ev***	*denárc**em***	*denárc**e***	*denárc**ih***	*denárc**i***		
				*grobôv**i***	*grobôv***	*grobôv**om***	*grobôv**e***	*grobôv**ih***	*grobôv**i***	*grobôv**a***	*grobôv**oma***
				*abstrákt**a*** (s)	*abstrákt*	*abstrákt**om***	*abstrákt**a***	*abstrákt**ih***	*abstrákt**i***		

Opomba: Nemi *e* se v tujkah ohrani, če določa izgovor predhodnega soglasnika (*bridge* [*bríd̄ž-*] *bridgea* [*brídža*] proti *brumaire* [*brimêr*] *brumaira* [*brimêra*]).

[6] Pri samostalnikih, ki poznajo podspol človeškosti oziroma živosti, je tožilnik ednine enak rodilniku, pri drugih pa imenovalniku.
[7] Za *c, j, č, ž, š, dž* se v končnici *o* premenjuje z *e*.

Slika 2: Primer sheme za pregibanje samostalnikov po prvi moški sklanjatvi (z opombami).

Pri tem na določitev podtipa vplivajo tudi morebitne dodatne oblike, navedene v zaglavju, ni pa nujno, saj so te lahko navedene kot posebnosti posamezne besedne vrste, ki ne določajo nadaljnjih oblik, ali pa so to zgolj oblike, pri katerih bi po leksikografovi presoji lahko nastal dvom (glej Uvod v SSKJ2: §184–185), pri čemer v uvodu ni podrobneje pojasnjeno, kako naj uporabniki razlikujejo med posameznimi možnostmi. Prav tako pa imajo v zaglavju več oblik navedene tiste iztočnice, ki jih ni mogoče pregibati glede na tipe v shemah, kot prikazuje primer na Sliki 3.



òn ôna -o stil. -ó zaim., ed. m. njêga, njêmu, njêga, njêm, njím, enklitično rod., tož. ga, daj. mu, enklitični tož. za enozložnimi predlogi -nj oziroma -enj [ənj] , če se predlog končuje na soglasnik; ž. njé, njêj tudi njèj tudi njì, njó, njêj tudi njèj tudi njì, njó, enklitično rod. je, daj. ji, tož. jo, enklitični tož. za enozložnimi predlogi -njo; s. kakor m., le tož. ônо stil. onó tudi njêga; mn. m. ôni stil. oní, njìh, njìm, njìh in njé, njih, njími, enklitično rod., tož. jih, daj. jim, enklitični tož. za enozložnimi predlogi -nje; ž. ône stil. oné dalje kakor m.; s. ôna stil. oná dalje kakor m.; dv. m. ônadva tudi onádva stil. ôna, njíju tudi njìh tudi njìh dvéh stil. njú, njíma tudi njíma dvéma, njíju tudi njìh tudi njìh dvá stil. njú, njíju tudi njìh tudi njìh dvéh tudi njíma tudi njíma dvéma, njíma tudi njíma dvéma, enklitično rod., tož. ju in jih, daj. jima, enklitični tož. za enozložnimi predlogi -nju; ž. ônidve stil. onédve dalje kakor m., le tož. njíju tudi njìh tudi njìh dvé stil. njú; s. kakor ž. (ö ó)

Slika 3: Prikaz pregibanja v zaglavju zaimka *on* v spletni različici SSKJ2.

Kljub dejstvu, da se SSKJ2 vzpostavlja tudi kot referenčni oblikoslovni priročnik, se tako zdi njegova oblikoslovna informativnost precej okrnjena, saj razvezava v oblikoslovno-naglasne sheme predstavlja velik uporabniški izziv, ki za uspešno reševanje oblikoslovne zadrege zahteva hkratno upoštevanje podatkov v konkretnem geslu in splošnejših, metajezikovno zahtevnih, navodil iz uvoda.[7] Čeprav je tak sistem podajanja oblikoslovnih informacij razumljiv z vidika omejitev tiskane zasnove SSKJ, je glede na spremenjene okoliščine izdaje SSKJ2 manj sprejemljiv, zlasti ob dejstvu, da je bilo že pri načrtovanju njegove zasnove izpostavljeno, da je razvezava oblikoslovno-naglasnih in tonemskih shem zahtevna tudi z vidika strokovnih uporabnikov (Perdih 2008: 18, 136, 142–143).

4 LEKSIKON BESEDNIH OBLIK KOT DEL DIGITALNO ZASNOVANE SLOVARSKE BAZE

Da digitalna zasnova slovarja omogoča drugačen način prikazovanja oblikoslovnih informacij, ugotavljajo tudi avtorji Predloga za izdelavo slovarja sodobnega slovenskega jezika (Krek et al. 2013), ki beleženje oblikoslovnih informacij predvidevajo znotraj ločene podatkovne baze, nadgrajenega leksikona besednih oblik Sloleks (glej prispevek Dobrovoljc et al. 2015), njihovo prikazovanje v spletnem slovarju pa znotraj ločenega zavihka *Oblika*. Zelo podobno rešitev predlagajo tudi avtorji Osnutka koncepta novega slovarja slovenskega knjižnega jezika (NSSKJ; Gliha Komac et al. 2015), ki kot enega izmed dveh glavnih gradnikov slovarske baze napovedujejo lematsko bazo s podatki o izrazni podobi slovarskih iztočnic, ki bi se v spletni različici slovarja prikazovali znotraj razdelka *Izgovor in oblike*.

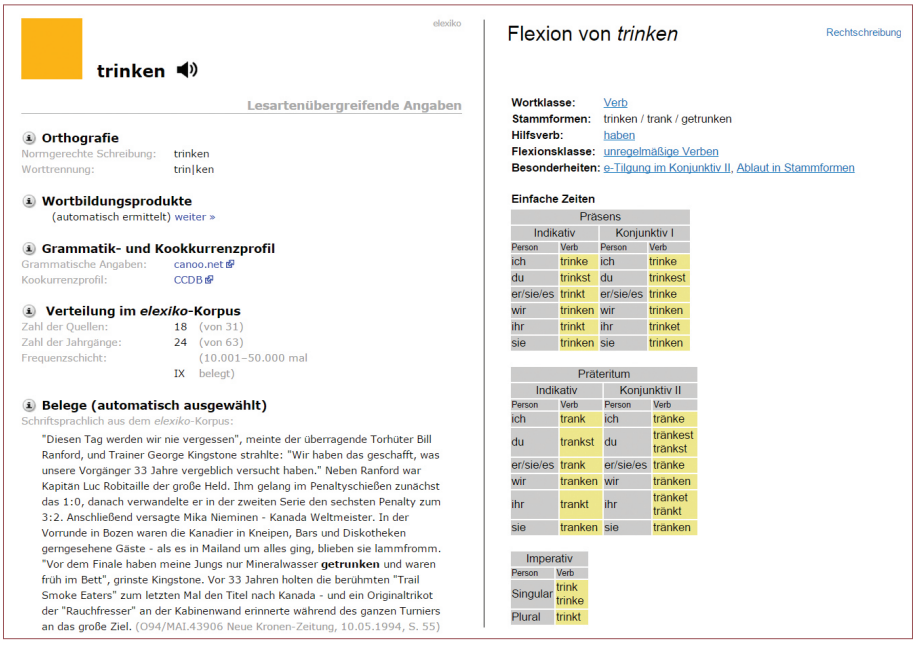
Usklajeno predlagana integracija oblikoslovnega leksikona kot samostojne podatkovne zbirke prinaša številne prednosti z vidika oblikoslovne informativnosti slovarja in splošne uporabniške izkušnje njegovih uporabnikov (4.1), a pri tem odpira nova vprašanja vsebinsko-tehničnega razmerja med leksikonsko in slovarsko iztočnico (4.2) ter razmerja med podatki v bazi in slovarju (4.3).

4.1 Leksikon besednih oblik kot vir in usmerjevalnik slovarskih informacij

Strojno berljiva oblikoslovna podatkovna zbirka, kot je leksikon besednih oblik, je poleg uporabnosti v jezikovnotehnoloških aplikacijah za označevanje

7 Zahtevano upoštevanje navodil v uvodu je še zlasti problematično z vidika šolskih uporabnikov in tujih govorcev, ki se jezika šele učijo, saj ti bistveno pogosteje potrebujejo tudi informacije o sistemsko razvezanih oblikah in modifikacijah, ki so za razliko od posebnosti opisane zgolj v shemah v uvodu.

korpusnih besedil (glej prispevek Erjavec et al. 2015a) v kontekstu povezovanja z digitalno zasnovano slovarsko bazo v prvi vrsti namenjena prikazovanju izraznih lastnosti slovarskih iztočnic, kot so podatki o njeni besedni vrsti in drugih oblikoslovnih lastnostih ter njeni pregibni, izgovorni in besedotvorni paradigmi. Glede na prakso primerljivih tujih spletnih slovarjev je paradigmatske informacije smiselno prikazovati v celoti, brez krajšav, in sicer bodisi prek povezave na zunanji oblikoslovni vir (kot denimo pri islandskem slovarskem portalu ISLEX, leksikalni podatkovni bazi za francoščino BFL ali nemškem jezikovnem portalu elexiko, kot prikazuje Slika 4), bodisi že v samem geselskem članku, pri čemer je vsebina pri slovarjih morfološko manj bogatih jezikov običajno navedena že na prvi ravni, ob iztočnici (npr. pri slovarskem portalu Collins ali splošnem nizozemskem slovarju ANW), pri drugih pa ob kliku na dodatni gumb ali zavihek (npr. pri španskem slovarju Daele ali Velikem slovarju poljskega jezika, kot prikazuje Slika 5).



Slika 4: Prikaz pregibne paradigme na spletnem portalu *Elexiko* s povezavo na oblikoslovni leksikon *Canoo*.

Slika 5: Prikaz pregibne paradigme na spletnem portalu Velikega slovarja poljskega jezika znotraj zavihka *Odmiana* (Oblika).

Ob predvideni kategorizaciji izrazne variantnosti in z njo povezane normativne kvalifikacije iztočnic in njenih oblik (glej prispevek Dobrovoljc et al. 2015) lahko leksikon besednih oblik obenem deluje tudi kot sidrišče normativnih informacij o morebitnih pravopisnih, pravorečnih, oblikoslovnih, besedotvornih, skladenjskih ali drugih jezikovnih zadregah, povezanih s slovarsko iztočnico. Ob predpostavki, da bo v kontekstu informativno-normativne zasnove bodočega slovarja za vsako izmed slovarsko relevantnih jezikovnih zadreg oblikovano daljše univerzalno pojasnilo za prikaz pri vseh iztočnicah z enakim tipom zadrege (glej npr. zavihek *Norma* v Krek et al. 2013b: 41 in Popič 2015), vsakokratni priklic pojasnila k iztočnici usmerja prav podatek o pripisani kategoriji v leksikonu besednih oblik, na podlagi katere se lahko na ustreznih mestih slovarskega sestavka (denimo ob iztočnici ali enem izmed njenih variantnih zapisov, ob posamični obliki ali izgovoru ipd.) samodejno prikazujejo tudi različni normativni kvalifikatorji oz. opozorila, ki uporabnika opozarjajo na relevantne posebnosti in ga usmerjajo k njihovemu pojasnilu.

V okviru povezovanja leksikona besednih oblik s slovarsko bazo pa ta ni zgolj podatkovni vir oblikoslovnih oz. normativnih informacij o posameznih iztočnicah, temveč deluje tudi kot usmerjevalnik prikazovanja drugih tipov slovarskih

informacij. Med najpomembnejšimi prednostmi je zagotovo možnost iskanja po pregibnih oblikah, ki v primerjavi s spletno različico SSKJ2 ali slovarskim portalom Fran uporabnikom omogoča intuitivnejše oblikovanje iskalnih pogojev, ne da bi ti pri tem morali poznati načela besednovrstne kategorizacije in lematizacije iztočnic ter pravopisnih pravil njihovega zapisovanja.[8] Na podoben način lahko leksikon pripomore tudi k razumljivosti slovarskih razlag in ponazoritev, saj uporabnika s klikom na posamezno nerazumljivo besedno obliko v slovarskem sestavku samodejno poveže z razlago ustrezne slovarske iztočnice (prim. npr. povezljivost razlagalnih besed na slovarskih portalih Wiktionary in TheFreeDictionary).

Poleg učinkovitejšega dostopanja do slovarskih informacij pa leksikon besednih oblik nenazadnje omogoča tudi njihovo povezljivost z zunanjimi besedilnimi viri in orodji. Tako je po vzoru spletnega portala Sloleks[9] preko leksikonskega podatka o lemi, zapisu in oblikoskladenjski oznaki posamezno obliko mogoče povezati s konkretnimi korpusnimi konkordancami (primeri rabe oblike v kontekstu), podatek o fonetični transkripciji pa omogoča samodejno sintezo izgovora prikazanih oblik na eni strani in samodejno prepoznavo glasovnih iskalnih poizvedb na drugi.

4.2 Razmerje med leksikonskim podatkom in slovarsko informacijo

Ob dejstvu, da beleženje oblikoslovnih informacij v obliki samostojne baze prinaša številne vsebinske in tehnične prednosti, je treba pri načrtovanju načina njihovega prikazovanja jasno vzpostaviti ločnico med izhodiščnimi podatki v leksikonski bazi na eni strani in slovarskemu uporabniku prikazanimi informacijami na drugi. Ena najpomembnejših prednosti hierarhično strukturirane strojno berljive podatkovne zbirke je namreč prav dejstvo, da omogoča dinamično prilagajanje načina prikazovanja informacij tipu priročnika in specifičnim potrebam njegovih uporabnikov, s čimer nimamo v mislih zgolj oblikovnih ali tehničnih rešitev, temveč tudi nabor in vsebino posredovanih informacij, med katerimi se kot potencialno najbolj problematična kažejo vprašanja vključevanja podatkov o nestandardnem, vključevanja podatkov o izgovoru in načina prikaza slovničnih informacij.

Čeprav poleg standardnih nestandardne pregibne variante navaja tudi večina obstoječih slovarskih priročnikov za slovenščino, so te običajno omejene le na nekaj ponavljajočih se pravopisno ali oblikoslovno težavnih posebnosti, kot

8 Raziskava dnevniških datotek iskanj po danskem spletnem slovarju Den Danske Netordbor (Bergenholtz in Johnsen 2005: 127–133) denimo kaže, da so med 19,5 % neuspelih poizvedb najpogostejše deležniške in velelniške oblike glagolov, zatipkane besede, napačni zapisi pod vplivom izgovora ter napačni zapisi skupaj ali narazen.

9 http://www.slovenscina.eu/sloleks

je denimo sklanjanje samostalnikov *otrok, mati, hči, gospa* ipd. V leksikonu besednih oblik, ki bi temeljil na izčrpnem opisu izraznih lastnosti slovarskih iztočnic, kot se kažejo v dejanski jezikovni rabi, lahko poleg tovrstnih posebnosti pričakujemo tudi pogoste sistemske nestandardne variantne paradigme in modifikacije, kot so nestandardne podaljšave, premene ipd. Izkušnje pri vizualizaciji leksikona besednih oblik Sloleks, ki nekaj takih primerov že vsebuje, kažejo, da je za razliko od standardnih paradigem, kjer so prikazane vse oblike paradigmatskega vzorca ne glede na pogostost v rabi, pri nestandardnih variantnih paradigmah smiselno prikazovati le tiste oblike, ki se v rabi tudi dejansko pojavljajo. Pri tem je seveda ena prvih nalog evalvacije uporabniških navad raziskati, kakšna je frekvenčna meja, pod katero prikazovanje nestandardnega v rabi nima več informativno-izobraževalne vloge, temveč je tudi ob premišljeni oblikovni rešitvi za uporabnika moteče.

Podobno velja tudi za prikazovanje informacij o izgovoru, saj lahko glede na količino v slovenščini pogostih naglasnih dvojnic predvidimo zelo obsežne izgovorne paradigme, če tem dodamo še variantnost v izgovoru posamičnih glasov ali več različnih načinov zapisa (neonaglašene ali onaglašene oblike, standardizirana ali druga izbrana fonetična abeceda), pa kombinatorika prikazanih oblik hitro postane neobvladljiva. Zato je pri načrtovanju prikazovanja izgovornih informacij smiselno razmišljati o prioritizaciji izgovornih informacij, denimo privzetega pripisa onaglašene oblike in fonetične transkripcije na ravni iztočnice, medtem ko so lahko izgovorni podatki o onaglašeni ali fonetično transkribirani paradigmi, ki jih tuji slovarji praviloma ne prikazujejo, prikazani ugnezdeno pod neonaglašenim zapisom oblike ali s klikom na dodatno polje oz. zavihek.

Podobno se je treba ločnice med leksikonskim podatkom in slovarsko informacijo zavedati tudi pri prikazovanju slovničnih informacij. Formalna slovnica v leksikonu besednih oblik, ki je prilagojena potrebam in omejitvam strojnega procesiranja slovenščine, namreč ni nujno enaka slovarski oz. priročniški slovnici. Poleg metajezikovnih terminoloških vprašanj, kot je preimenovanje laični javnosti manj razumljivih slovničnih lastnosti (npr. nedoločnost, dvovidskost ipd.) in siceršnji delež njihovega prikazovanja, to zajema tudi povsem temeljne odločitve o načinu določanja in razvrščanja besednih vrst in drugih oblikoslovnih lastnosti, npr. o morebitnem združevanju deležij z izvornimi glagoli, elativov z drugimi stopnjami ipd. Razlikovanje med formalno in priročniško slovnico samo po sebi ni problematično, pomembno pa je, da so preslikave med eno in drugo sistematizirane in ustrezno dokumentirane, saj le jasno opredeljena razmerja med podatki v izhodiščni bazi in načinom njihovega vključevanja v druge leksikalne zbirke omogočajo dolgoročno povezljivost jezikovnih virov.

4.3 Razmerje med leksikonsko in slovarsko iztočnico

Leksikon besednih oblik je privzeto namenjen beleženju oblikoslovnih ter deloma besedotvornih in normativnih podatkov o besedišču, ne pa pomenskih, kar pomeni, da so leme z identičnimi oblikovnimi, izgovornimi in slovničnimi lastnostmi opisane kot del iste leksikonske enote, ne glede na razliko v njihovem pomenu. Tako so v leksikonu po eni strani združene homonimne leme, kot sta para *bor* (drevo) in *bor* (kemijski element) ali *početi* (začeti) in *početi* (opravljati), po drugi strani pa ločene pomensko potencialno prekrivne leme, kadar imajo te drugačno izrazno podobo, denimo *volivec* in *volilec, posebej* in *posebaj*, *zvedeti* in *izvedeti*. Pri razmislekih o načinu povezovanja leksikona besednih oblik in slovarskih iztočnic je torej treba upoštevati, da glede na različna načela zasnove in namembnosti obeh podatkovnih baz razmerje med leksikonsko in slovarsko iztočnico ni nujno simetrično, niti ne statično, saj je odvisno predvsem od dogovorno izbranih meril za opredelitev iztočnic v posamezni leksikalni zbirki.

Pri merilih za opredeljevanje slovarskih iztočnic, ki vplivajo na način povezovanja leksikonske in slovarske baze, je tako med ključnimi predvsem vprašanje formalnega razlikovanja med homonimijo in večpomenskostjo, torej vprašanje, katere izrazne lastnosti pomensko različnih enakopisnih lem so pogoj za njihovo obravnavo v obliki samostojnih slovarskih iztočnic (homonimija) ali njihovo obravnavo znotraj iste slovarske iztočnice (večpomenskost). Kot ugotavlja Gantar (2015), se tako teoretsko kot uporabniško motivirani pristopi k obravnavi homonimije in večpomenskosti običajno strinjajo, da so ne glede na morebitno pomensko ali etimološko sorodnost kot ločene slovarske iztočnice obravnavane enakopisnice z različno besedno vrsto (samostalnik in prislov *naglas*, samostalnik in pridevnik *žužkojed*), različnimi slovničnimi lastnostmi (ženski in moški samostalnik *prst*, moški in srednji samostalnik *čelo*), celotno pregibno paradigmo (nedovršna glagola *vesti: vezem* in *vesti: vedem*, dovršna glagola *postati: postanem* in *postati: postojim*) ali njenim izgovorom (*molíti: molím* in *móliti: mólim*, *partíja: partíje* in *pártija: pártije*), kar se sklada tudi s stanjem v leksikonu besednih oblik, ki pregibanje tovrstnih enot opisuje znotraj samostojnih leksikonskih iztočnic (razmerje med obema bazama je v takšnih primerih torej simetrično).

Sklicevanje na formalna merila razlikovanja je manj dorečeno, če kot izrazne lastnosti homonimnih lem upoštevamo tudi (ne)prekrivnosti njihovih besedotvorno povezanih oblik (npr. *vila*, kjer se besedotvorna povezava *vilinski* povezuje samo s pomenom 'mitološko bitje' in ne 'razkošna hiša'), dela paradigme (npr. *bučen*, kjer se stopnjevane oblike povezujejo samo s pomenom 'glasen' in ne 'nanašajoč se na bučo', ali *lisica*, kjer se pomen 'naprava za zaklepanje' povezuje samo z množinsko paradigmo) ali posameznih oblik (npr. *tenor*, kjer se oblika v

tožilniku ednina razlikuje glede na podspol živosti). Glede na rabo leksikona v raznih jezikovnotehnoloških aplikacijah, pri katerih ni mogoče pričakovati pomenskega razdvoumljanja enakih oblik z enakimi slovničnimi lastnostmi (npr. oblik *bučnega, lisic* ali *tenorja* v različnih pomenih), je v leksikonu besednih oblik praviloma upoštevano načelo maksimalne paradigme, torej združevanja prekrivnih paradigem, tudi če se posamezni pomeni uresničujejo samo v posamičnih slovničnih kategorijah ali oblikah. Ne glede na odločitev glede obravnave določenega pomena v obliki samostojne ali skupne slovarske iztočnice, je pri tovrstnih primerih razmerje med leksikonsko in slovarsko iztočnico torej nesimetrično, saj se s posamezno slovarsko iztočnico oz. pomenom povezuje samo del neke leksikonske enote (npr. *lisica*, *bučen*, *tenor*), kar je treba upoštevati tako pri tehničnih kot oblikovnih rešitvah slovarskega portala.[10]

Če smo v prejšnjem odstavku načeli vprašanje povezovanja (dela) ene leksikonske enote z več slovarskimi iztočnicami oz. pomeni, je po drugi strani enako relevantno tudi vprašanje povezovanja ene slovarske iztočnice z več leksikonskimi. Tipičen primer so variantni zapisi, npr. *žiroračun* in *žiro račun*, *eventuelno* in *eventualno*, *volivec* in *volilec*, ki so v leksikonu besednih oblik obravnavani kot samostojne iztočnice, v primeru ugotovljene pomenske prekrivnosti na podlagi prekrivnega besedilnega okolja pa bi v slovarski bazi lahko bile združene pod eno skupno iztočnico z več variantnimi zapisi in z njimi povezanimi oblikoslovnimi paradigmami. Podobno velja tudi za potencialno pomensko prekrivne pare enakopisnic z variantnimi, leksikonsko razločevalnimi, slovničnimi lastnostmi, kot npr. pri samostalnikih *činčila*, *sluz* in *nadlaket*, ki se v rabi pojavljajo kot samostalniki tako moškega kot ženskega spola, ali *finale*, ki se pojavlja kot samostalnik tako moškega kot srednjega spola. Tudi če bi bili glede na izbrana merila določanja slovarskih iztočnic tovrstni primeri obravnavani v obliki samostojnih iztočnic (torej simetrično z leksikonom besednih oblik), pa je priklic več ločenih leksikonskih enot znotraj ene same slovarske iztočnice neizogiben pri iztočnicah z raznospolskim pregibanjem, kot sta denimo samostalnika *oko* (v enem izmed pomenov v množini tudi kot *oči*) ali *ledvica* (v množini *ledvice* ali *ledvica*).

5 ZAKLJUČEK

Tako slovenska in tuja slovaropisna tradicija kot empirično utemeljene analize uporabniških potreb dokazujejo, da se s pomenskim opisom besedišča nepogrešljivo povezuje tudi opis njegovih izraznih lastnosti. Pri tem je pri načrtovanju

10 Enega izmed možnih odgovorov na vprašanje, kako prikazovati pomensko pogojene oblikoslovne specifike, denimo ponuja Veliki slovar poljskega jezika, ki enakopisnice ne glede na njihovo etimološko prekrivnost obravnava kot večpomenske iztočnice, pri čemer so slovnične, pomenske in skladenjske informacije, vključno s prikazom pregibne paradigme, uporabniku prikazane šele po izbiri posameznega pomena.

vključevanja oblikoslovnih informacij v bodoče slovarske opise slovenskega jezika smiselno prekiniti z dosedanjim načinom podajanja pregibnih in drugih oblikoslovnih informacij, kakršen je bil glede na omejitve prvotne knjižne zasnove vzpostavljen s prvo izdajo SSKJ in se je kljub spremenjenim družbenim in tehnološkim okoliščinam pri referenčnih slovarskih priročnikih ohranjal vse do danes, saj je njihova oblikoslovna informativnost izrazito okrnjena. S spletnim medijem in z njim povezano digitalno zasnovo bodočih slovarskih opisov je beleženje oblikoslovnih podatkov glede na morfološko bogatost slovenskega jezika smiselno načrtovati v obliki ločene, a s slovarsko bazo premišljeno usklajene podatkovne baze, ki omogoča njeno dolgoročno povezovanje tudi z drugimi jezikovnimi priročniki ali jezikovnotehnološkimi aplikacijami, hkrati pa tako v vsebinskem, oblikovnem in tehničnem smislu omogoča dinamično prilagajanje raznolikim uporabniškim potrebam.

Leksikon besednih oblik Sloleks in smernice njegovega razvoja

Kaja Dobrovoljc, Simon Krek in Tomaž Erjavec

Abstract

This paper presents Sloleks, the largest open-source machine-readable morphological lexicon of the Slovenian language to date. The development of Sloleks and the formal grammar behind it are first briefly presented, before a detailed presentation of the types and structure of the inflectional, derivational, grammatical and other information included in each lexical entry is provided, with special emphasis on its formal representation within the standardised XML LMF framework. Given that Sloleks is a strong candidate to be used in the compilation of a new dictionary of contemporary Slovenian, both as a source of morphological information as well as part of the language technologies behind it, the second part of the presentation explores the most important aspects of its future development, particularly the expansion of its entry list, the addition of pronunciation information, the normative categorisation of variants and a corpus-based re-evaluation of existing inflectional paradigms. An extensive usage-based open-source morphological lexicon of modern Slovenian, with a universally unified system of morphological description, will therefore have a long-term application for language technologies and other born-digital reference works for the Slovenian language.

Keywords: morphological lexicon, lexicon of inflected forms, machine readable dictionary, morphology, inflection, derivation, pronunciation, language standardisation

Ključne besede: oblikoslovni leksikon, leksikon besednih oblik, strojno berljivi slovar, oblikoslovje, pregibanje, besedotvorje, izgovor, jezikovna standardizacija

1 UVOD

Pri oblikoslovno bogatih jezikih, kot je slovenščina, je opis oblikoslovnih paradigem pri pregibnih besednih vrstah tradicionalno zelo pomemben. Že Bohoričeva slovnica (1584) skoraj polovico opisa namenja poglavju o pregibanju besed oz. etimologiji, kot to imenuje Bohorič. Oblikoslovne paradigme imajo podobno prominentno vlogo tudi v večini kasnejših slovenskih slovnic. Te se osredotočajo predvsem na sistemske vidike oblikoslovja, torej na oblikoslovne vzorce, ki jih konkretno ponazorijo z nekaj primeri, kar pomeni, da je eksplicitnih izpisov celotnih paradigem v slovnicah malo. Po drugi strani so v preddigitalni dobi slovarji, predvsem v obliki pravopisnih priročnikov, kasneje pa SSKJ, kot popisovalci besedišča tradicionalno vsebovali tudi podatke o pregibanju. V njih so morfološki opisi močno okrajšani, poleg iztočnice navadno omejeni na eno ali nekaj oblik paradigme, na podlagi katerih naj bi uporabniki sklepali na celotno oblikoslovno paradigmo. Tudi po prenosu tiskanih priročnikov v digitalno obliko so ti podatki ostali identični.

S pojavom računalnikov in razvojem področja procesiranja besedil v naravnih jezikih je bila kmalu izpostavljena potreba po dostopnosti strojno berljivih slovarjev in leksikonov besednih oblik (Atkins in Zampolli 1994). Za angleščino so bili prvi strojno berljivi slovarji za različne jezikovnotehnološke naloge izdelani že v šestdesetih letih prejšnjega stoletja (npr. Boguraev in Briscoe 1987), s splošno digitalizacijo jezikov v devetdesetih letih pa so začeli nastajati tudi oblikoslovni leksikoni za večino drugih evropskih jezikov.

Ker za računalniške potrebe ni dovolj, da je na voljo le vzorec ali nekaj oblik, so v teh leksikonih paradigme glede na razbremenjenost prostorskih omejitev tiskanega medija tipično izpisane v celoti in se nahajajo v formatu, ki je strojno berljiv. Oblikoslovni podatki, tradicionalno vsebovani v slovnicah in slovarjih in namenjeni uporabnikom knjižnih jezikovnih priročnikov, so torej z računalnikom kot novim »uporabnikom« dobili tudi novo področje aplikacije. Leksikoni morajo tako hkrati zadovoljiti jezikovnotehnološke potrebe v različnih računalniških aplikacijah, od črkovalnikov in slovničnih označevalnikov do razpoznavalnikov in sintetizatorjev govora, strojnih prevajalnikov ipd., po drugi strani pa je zaželeno, da so uporabni tudi kot samostojni, jezikovnim uporabnikom namenjeni oblikoslovni priročniki. Sodoben računalniški leksikon slovenskega jezika naj bi torej bil namenjen zadovoljevanju obeh potreb in zato organiziran drugače kot oblikoslovni podatki v tradicionalnih slovarjih in slovnicah ter tudi kot prvotni strojno berljivi leksikoni.

Sledenje tema dvema ciljema vsebinsko zasnovo leksikona postavlja pred dve nasprotujoči si tendenci: v jezikovnotehnoloških aplikacijah mora leksikon čim

bolj uspešno opredeljevati oblikoslovne lastnosti vseh besednih oblik, ki jih srečamo v realnih besedilih, vključno z govorjenimi besedili, in omogočiti preprosto strojno berljivost pripisanih podatkov. Pri tradicionalni slovarski rabi pa mora zagotoviti učinkovito podajanje pregibnih, izgovornih in besedotvornih informacij za človeškega uporabnika, vključno z normativnimi vidiki besedišča. V kontekstu rabe leksikona za potrebe izdelave bodočega slovarja sodobnega slovenskega jezika mora biti leksikon v tem smislu vsebinsko usklajen z obema poloma: po eni strani z oblikoslovnimi podatki v učnem korpusu, s pomočjo katerega se oblikoslovni označevalniki učijo strojno označevati besedilne korpuse, iz katerih pridobivamo podatke za slovar, po drugi strani pa mora biti leksikon usklajen z leksikografskimi podatki v slovarski bazi ali drugih povezanih podatkovnih zbirkah.

Glede zadovoljevanja tradicionalnih jezikovnopriročniških potreb je pri oblikovanju vsebine referenčnega oblikoslovnega leksikona za sodobni slovenski jezik ključna težava v tem, da so obstoječi referenčni jezikovni priročniki, torej slovnice (npr. Toporišič 2004), slovarji (npr. SSKJ2) in pravopisi (npr. SP 2001) glede obravnave oblikoslovnih podatkov neusklajeni, celo kontradiktorni (prim. Krek 2014), kar pomeni, da za izhodišče ni mogoče vzeti nobenega od omenjenih del, temveč je treba koncept vsebinsko oblikovati na novo. Ti priročniki v pretežni meri tudi niso nastajali na podlagi sodobnega gradiva, zato so razmeroma oddaljeni od jezikovne realnosti sodobne slovenščine, po kakršni poizvedujejo tako jezikovnopriročniški kot jezikovnotehnološki uporabniki.

Strojno berljivi leksikoni besednih oblik za slovenščino imajo razmeroma dolgo zgodovino. Na začetku devetdesetih let je podjetje Amebis začelo razvijati elektronski slovar slovenskega jezika ASES, ki vsebuje tudi eksplicirane oblikoslovne paradigme (Arhar in Holozan 2009). Baza tega slovarja oz. leksikona ni prosto dostopna, podatke, ki jih vsebuje, pa je mogoče najti v različnih izdelkih podjetja, kot so slovnični pregledovalnik Besana, strojni prevajalnik Presis, sistem za komunikacijo v naravnem jeziku itd. Kronološko gledano je bil prvi prosto dostopni računalniški leksikon za slovenski jezik izdelan v okviru projekta MULTEXT-East v devetdesetih letih in vsebuje prek 15.000 osnovnih oblik ali lem in njihove pregibne paradigme v tabelaričnem formatu (Erjavec et al. 1995).

V prvem desetletju tega stoletja so z razvojem govornih tehnologij, predvsem sinteze govora, postali pomembni tudi leksikoni, ki poleg oblikoslovnih podatkov vsebujejo informacije o izgovoru, denimo SIflex, SImlex (Rojc et al. 2002; Verdonik et al. 2002), LC-STAR (Verdonik et al. 2004; Verdonik in Rojc 2004), SI-PRON (Žganec Gros et al. 2006) itd. Precejšnjo težavo pri vseh omenjenih leksikonih s podatki o izgovoru predstavlja dejstvo, da niti eden ni prosto dostopen. Enako je z dostopnostjo računalniškega oblikoslovnega leksikona, ki je

bil v približno istem času izdelan na Inštitutu za slovenski jezik Frana Ramovša, vendar o njem ni na voljo nobenih podatkov razen navedbe o obstoju (Naglič et al. 2005: 36).

Nekoliko bolj specifičen je še leksikon, ki je na voljo v prosto dostopnem sistemu za strojno prevajanje Apertium in obsega nekaj nad 20.000 lem (Horvat in Vičič 2012; Vičič 2012). Čeprav ta v osnovi izhaja iz leksikona MULTEXT-East, je glede vsebine in formata nekoliko drugačen, ker je v precejšnji meri vezan na prevajalni sistem, zato ni uporaben kot splošni leksikon za slovenščino. V okviru projekta Sporazumevanje v slovenskem jeziku je nastal računalniški leksikon Sloleks (Dobrovoljc et al. 2013), ki ga tudi postavljamo v središče tega prispevka, saj glede na svoj obseg, odprto dostopnost in rabo v temeljnih jezikovnotehnoloških orodjih za slovenščino predstavlja smiselno izhodišče za nadaljnji razvoj referenčnega leksikona besednih oblik za slovenščino.

2 LEKSIKON BESEDNIH OBLIK SLOLEKS

V pričujočem razdelku podrobneje predstavimo vsebino leksikona besednih oblik Sloleks, format njegovega zapisa, nabor in organiziranost podatkov posamične leksikonske enote ter način njihovega prikazovanja v obstoječem spletnem vmesniku.

2.1 Vsebina leksikona

2.1.1 Geslovnik in paradigme

Leksikon besednih oblik Sloleks v trenutni različici (Dobrovoljc et al. 2013) vsebuje 100.805 gesel, pri čemer eno geslo ustreza eni lemi, njeni pregibni paradigmi in drugim oblikoslovnim podatkom. Nabor iztočnic oz. lem je bil izdelan na podlagi meril, določenih v specifikacijah za izdelavo leksikona besednih oblik (Erjavec et al. 2008), in sicer je bila v leksikon najprej zajeta večina lem ročno označenega učnega korpusa ssj500k (Krek et al. 2013a), vse leme leksikalno zamejenih besednih vrst (predlog, veznik, zaimek, členek) ter izbrani težji primeri iz oblikoslovja npr. tuja lastna imena, enakovidski glagoli s homonimnimi nedoločniki (npr. *stati*), moški samostalniki z živo in neživo obliko v tožilniku ednine (npr. *delfin*), težavni primeri z visoko stopnjo variantnosti (npr. *otrok*) itd. Preostala večina geselskih iztočnic je bila nato izbrana glede na pogostost rabe na podlagi seznama najpogostejših lem v takratnem referenčnem korpusu FidaPLUS v obsegu 620 milijonov besed (Arhar in Gorjanc 2007).

V drugi fazi izdelave leksikona so bile lemam pripisane pregibne oblike, s programom za polavtomatsko dodajanje oblikoslovnih paradigem, ki ga je razvilo podjetje Amebis d. o. o. za izdelavo podatkovne zbirke ASES (Arhar in Holozan 2009) in na njej temelječih orodij. Leksikon besednih oblik Sloleks vsebuje skoraj 2.800.000 pregibnih oblik, točno sestavo leksikona glede na število lem in oblik posameznih besednih vrst prikazujemo v Tabeli 1.

Tabela 1: Število lem in oblik v leksikonu besednih oblik Sloleks v1.2.

Besedna vrsta	Število lem	Število oblik
samostalniki	54.260	924.268
pridevniki	26.612	1.571.970
glagoli	10.242	260.826
prislovi	6.906	9.931
števniki	2.240	18.448
zaimki	169	6.182
predlogi	96	101
medmeti	85	85
okrajšave	70	70
členki	68	68
vezniki	54	54
večbesedne enote[1]	3	3
SKUPAJ	**100.805**	**2.792.006**

2.1.2 Sistem JOS

Slovnične informacije v leksikonu besednih oblik temeljijo na oblikoskladenjskih specifikacijah, razvitih v okviru projekta Jezikoslovno označevanje slovenščine (Erjavec in Krek 2008; JOS), z namenom označevanja besednih pojavnic v slovenskih besedilih. Sistem JOS je rezultat daljše tradicije razvoja računalniških slovnic za slovenščino, saj je usklajen s specifikacijami projektov MULTEXT (Ide in Véronis 1994) in nato MULTEXT-East, pri čemer so specifikacije MULTEXT-East 4.0 (Erjavec 2012), ki med skupno 12 jeziki pokrivajo večino slovanskih jezikov, za slovenščino identične specifikacijam JOS.

Specifikacije JOS definirajo dvanajst besednih vrst: samostalnik, pridevnik, glagol, prislov, zaimek, števnik, predlog, veznik, členek, medmet, okrajšavo in neuvrščeno, pri čemer se zadnja v leksikonu ne uporablja. Z izjemo členkov, medmetov

1 Večbesedne enote so bile v leksikon vključene kot del poskusnih gesel na portalu Slogovni priročnik.

in okrajšav so večini besednih vrst pripisane dodatne oblikoskladenjske lastnosti, toda ni nujno, da so vsem besednim oblikam posamezne besedne vrste vedno pripisane vse možne lastnosti. Nabor možnih kombinacij slovničnih lastnosti je opredeljen v obliki vnaprej definiranega seznama[2] 1.902 kombinacij besednovrstnih kategorij, oblikoskladenjskih lastnosti in njihovih vrednosti, navodila za njihovo pripisovanje pojavnicam v besedilih pa so podrobneje opisana v navodilih za označevanje učnega korpusa (Holozan et al. 2008).

Kot ponazarjajo Erjavec et al. (2015), so bile oblikoskladenjske specifikacije sistema JOS razvite predvsem za potrebe strojnega označevanja besedil, zato zaradi omejenih možnosti označevalnih orodij na nekaterih mestih odstopajo od uveljavljenih slovnic slovenskega jezika (Ledinek 2014: 34–48). Pri določanju besedne vrste se tako upošteva predvsem oblika besede, manj pa njena skladenjska vloga v besedilu. Tipičen primer so denimo deležniške oblike na -n, -t ali -č, ki so ne glede na skladenjsko vlogo vedno označene kot deležniški pridevniki, saj na trenutni stopnji razvoja oblikoskladenjskih označevalnikov ni mogoče pričakovati dovolj zanesljivega razdvoumljanja njihove prilastkovne ali povedkovne vloge. Podobne poenostavitve so bile upoštevane tudi pri določanju posameznih oblikoskladenjskih lastnosti, kjer je denimo lastnost oseba pripisana vsem glagolskim pojavnicam v sedanjiku (tudi če so te brezosebne, npr. *dežuje*), (ne)določnost pa vsem pridevnikom, čeprav svojilni pridevniki te lastnosti ne razlikujejo.

S sistemom JOS so bila implicitno, skozi proces označevanja učnega korpusa in izdelave leksikona besednih oblik, opredeljena tudi načela določanja osnovnih oblik (lem) pojavnic pregibnih besednih vrst. Ta se večinoma skladajo z lematizacijskimi načeli v obstoječih priročnikih za slovenščino, npr. imenovalniška oblika ednine pri samostalnikih, nedoločniška oblika pri glagolih, nestopnjevana nedoločna oblika moškega spola ednine pri pridevnikih in besednih številnikih ter nestopnjevana oblika pri prislovih, z nekaj sistemskimi izjemami.[3] Posebnost so zaimki, pri katerih je lema odvisna od vrste zaimka in njegovih lastnosti (npr. leme *vame, zame, čezme* itd. za navezne osebne zaimke, ki se pregibajo po številu, osebi oz. spolu, ali lema *se* za povratne osebne zaimke *sebe/se, sebi/se, sabo/seboj*).

2.2 Zapis

Pri zasnovi leksikona kot referenčne zbirke oblikoslovnih, besedotvornih in drugih sorodnih podatkov o slovenskem jeziku je poleg proste dostopnosti kot

2 http://nl.ijs.si/jos/msd/html-sl/msd.index.msds.html

3 Pri množinskih samostalnikih je kot lema denimo izbrana oblika v imenovalniku množine (*alimenti*) oz. edina možna oblika (*poštev*). Pri prislovu so primerniške in presežniške oblike *bolj*, *manj*, *prej*, *raje*, *več*, *večkrat* oz. *najbolj*, *najmanj*, *najprej*, *najraje*, *največ* oz. *največkrat* zaradi specifičnih skladenjskih vlog osamosvojene v svojo lemo.

predpogoja za splošno rabo pomembno tudi upoštevanje standardov, ki omogočajo fleksibilno strukturiranje vsebine in mednarodno primerljivost podatkov.[4] Leksikon besednih oblik Sloleks je tako izhodiščno zapisan v označevalnem jeziku XML v shemi Lexical Markup Language (LMF), mednarodnem standardu za zapis strojno berljivih leksikalnih podatkovnih zbirk za potrebe procesiranja naravnih jezikov (ISO 24613:2008), ki je bil razvit z namenom vzpostavitve skupnega modela za izdelavo in uporabo eno- in večjezičnih leksikalnih virov, upravljanja izmenjave podatkov med dvema ali več viri ter združevanja večjega števila posameznih virov v obsežnejše globalne elektronske vire (Francopoulo et al. 2006: 1).

Format LMF sestavljajo t. i. jedrni sklop (angl. *core package*) in njegove razširitve (angl. *extensions*). Jedrni sklop predstavlja strukturno ogrodje, ki opisuje nabor in hierarhijo univerzalnih podatkov v leksikalni podatkovni zbirki, kot so informacija o jeziku, imenu in dostopnosti vira, ter osnovni strukturi leksikalnih enot, razširitve jedrnega sklopa pa nato določajo način kombiniranja strukturnih gradnikov (iz jedrnega sklopa) z drugimi elementi (iz razširitev) za potrebe specifičnega leksikalnega vira, kot je denimo oblikoslovni leksikon.[5]

Prilagoditev formata LMF za zapis oblikoslovnih podatkov morfološko bogatih jezikov, kakršna je bila upoštevana tudi pri izdelavi leksikona besednih oblik Sloleks, je podrobneje opisana v Krek in Erjavec (2009), celoten nabor pričakovanih elementov, atributov in možnih vrednosti ter njihova hierarhična razporeditev pa sta opisana v pripadajoči shemi DTD (*Document Type Definition*), ki je namenjena validaciji podatkov v bazi, torej preverjanju skladnosti njene strukture in vsebine z izhodiščnimi načeli.

2.3 Struktura leksikonske enote

Osnovno enoto leksikona besednih oblik, v kateri so strukturirani podatki enega gesla, imenujemo leksikonska enota. Ena leksikonska enota ustreza eni osnovni obliki oz. lemi in njeni oblikoslovni paradigmi, torej naboru ene ali več pregibnih oblik s pripadajočimi slovničnimi lastnostmi. Vsaka leksikonska enota obvezno vsebuje podatek o lemi, besedni vrsti in vsaj eni besedni obliki, glede na besedno vrsto in druge značilnosti leme pa je tej lahko pripisano še poljubno število dodatnih pregibnih oblik ter slovničnih in drugih lastnosti. V nadaljevanju na kratko opišemo nabor vseh tipov in hierarhična razmerja oblikoslovnih podatkov v leksikonu besednih oblik Sloleks in ponazorimo njihov zapis v formatu XML LMF.

4 Prvi prosto dostopni leksikoni so bili v shranjeni v tabelaričnem formatu, ki ni najboljši format za zapis možnosti, da npr. obstaja več variantnih besednih oblik z več izgovori, ti pa so denimo na kompleksen način povezane z drugimi oblikami.

5 Razširitve pri tem določajo predvsem pričakovani nabor elementov v določenem tipu vira, njihovo število in hierarhično urejenost, ne pa njihove semantične vsebine, saj so standardizirana poimenovanja jezikoslovnih kategorij, kot so poimenovanja besednih vrst, lastnosti in vrednosti, določena v registru podatkovnih kategorij ISOcat (http://www.isocat.org/).

2.3.1 Iztočnica

Iztočnica oz. ključ leksikonske enote je opredeljena kot unikatni identifikator, po katerem se leksikonske enote ločijo med seboj, saj med njimi ni mogoče ločevati zgolj na podlagi iztočniške leme, ki se lahko v enakem zapisu pojavlja v več leksikonskih enotah različnih besednih vrst (npr. prislov in členek *ravno,* prislov in samostalnik *stran,* prislov in pridevnik *spet*) ali znotraj iste besedne vrste (npr. dovršni in nedovršni glagol *zlagati*, deležniški in splošni pridevnik *poročen*, ženski in moški samostalnik *prst*). Čeprav je iztočnica namenjena predvsem strojni obdelavi podatkov, ne pa neposrednemu prikazovanju uporabnikom, je v leksikonu zasnovana tako, da lahko iz nje razberemo podatek o besedni vrsti in lemi (govoreča šifra), npr. S_avtomobil. V primeru, da se znotraj iste besedne vrste pojavlja več enakih lem, je temu zapisu dodana še zaporedna številka, npr. G_vesti_1 (*vesti: vezem*) ali G_vesti_2 (*vesti: vedem*).[6]

```
<LexicalEntry id="LE_ebc318126ea71205d05cd0ce85f86362">
          <feat att="ključ" val="R_pazljivo"/>
```

Slika 1: Zapis iztočnice prislova *pazljivo* v formatu XML LMF.

2.4.2 Lema

Osrednji gradnik leksikonske enote, na katerega se pripenjajo vsi drugi oblikoslovni podatki, je lema. Lema predstavlja neonaglašeno osnovno, kanonično oz. citatno besedno obliko, pod katero so združene vse druge oblike z enakimi leksikalnimi in paradigmatskimi lastnostmi, običajno pa tudi z enakim pomenom oz. pomeni. Določanje osnovne oblike v leksikonu besednih oblik Sloleks sledi lematizacijskim načelom sistema JOS, ki so bila upoštevana tudi pri lematizaciji učnega korpusa (Holozan et al. 2008) ter razvoju programa za samodejno lematizacijo in oblikoskladenjsko označevanje slovenskih besedil (Grčar et al. 2012).

```
<Lemma>
     <feat att="zapis oblike" val="pazljivo"/>
</Lemma>
```

Slika 2: Zapis leme prislova *pazljivo* v formatu XML LMF.

6 Posebnost so moški in ženski priimki, ki se v leksikonu vedno pojavljajo v paru in jim je zato namesto številke pripisana informacija o spolu, npr. S_Novak_m (Novak: Novaka) in S_Novak_ž (Novak: Novak). Kadar se pri priimkih pojavlja prekrivni samostalnik z enako lemo in spolom, se tudi tem iztočnicam doda zaporedna številka, npr. S_Pavlica_ž_1 (za nesklonljivi ženski priimek Pavlica) in S_Pavlica_ž_2 (za sklonljivo žensko ime Pavlica).

2.4.3 Besedna vrsta in leksikalne lastnosti

Obvezna slovnična lastnost vsake leksikonske enote in temeljna uvrščevalna lastnost leme je podatek o besedni vrsti, poleg tega pa je večini lem na prvi ravni strukture leksikonske enote pripisana še ena ali več leksikalnih lastnosti. Leksikalne lastnosti so tiste slovnične lastnosti, ki veljajo za vse oblike v paradigmi in jih pripišemo na ravni leme, npr. vrsta (občno ali lastno ime) in spol pri samostalnikih, vrsta (glavni ali pomožni) in vid (dovršni, nedovršni ali dvovidski) pri glagolih, sklon pri predlogih itd. Po vzoru formalnih slovničnih opisov so leksikalne in druge slovnične lastnosti zabeležene v obliki parov lastnosti (npr. *spol* pri samostalnikih) in njihovih vrednosti (npr. *moški*, *ženski* ali *srednji*).

```
<feat att="besedna_vrsta" val="prislov"/>
<feat att="vrsta" val="splošni"/>
```

Slika 3: Zapis leksikalnih lastnosti (vrste) prislova *pazljivo* v formatu XML LMF.

2.4.4 Pregibna paradigma

Splošnim podatkom o identifikatorjih, lemi in leksikalnih lastnostih sledi izpis pregibne paradigme, ki jo sestavljajo ena ali več pregibnih oblik, njihove partikularne oblikoslovne lastnosti, podatki o pogostosti v referenčnem korpusu in normativne lastnosti morebitnih variantnih oblik.

2.4.4.1 Pregibne oblike

Vsaka leksikonska enota ima v paradigmi[7] vsaj eno obliko. V primeru nepregibnih besednih vrst je ta oblika običajno samo ena, v primeru pregibnih vrst pa je teh oblik več, njihov obseg pa je odvisen od besedne vrste, leksikalnih lastnosti in stopnje variantnosti v jezikovni rabi. Med pregibnimi besednimi vrstami so tako najkrajše paradigme stopnjevanih prislovov in nekaterih zaimkov, najdaljše pa so paradigme pridevnikov, ki se pregibajo po spolu, stopnji, številu, sklonu in določnosti ter v povprečju obsegajo 59 različnih oblik (prim. Tabela 1).

2.4.4.2 Pregibne lastnosti

Posameznim oblikam so v paradigmi pripisane tudi pregibne slovnične lastnosti. V primerjavi z leksikalnimi lastnostmi so pregibne lastnosti tiste, po katerih

7 Z izrazom pregibna paradigma označujemo vse pregibne oblike leme, kot jih določa sistem JOS, ne glede na to, ali so rezultat oblikospreminjevalnih (npr. sklanjanje) ali oblikotvornih (npr. stopnjevanje) procesov.

se oblike v paradigmi določene leme z določenimi leksikalnimi lastnostmi ločijo med seboj, zato jih pripišemo na ravni (abstrahiranih) slovničnih oblik, npr. sklon, število in živost pri samostalnikih; stopnja pri prislovih; oblika, oseba, število, spol in nikalnost pri glagolih itd. Nabor pregibnih lastnosti v leksikonu besednih oblik Sloleks temelji na sistemu JOS, pri čemer ni nujno, da se vse pregibne lastnosti neke besedne vrste uresničujejo pri vseh njenih lemah, temveč je njihov dejanski nabor odvisen od same leme ali njenih leksikalnih lastnosti.
Na ravni pregibnih lastnosti je vključen tudi podatek o sintetični preslikavi vseh slovničnih lastnosti besedne oblike v t. i. oblikoskladenjsko oznako, kakršna se uporablja pri strojnem slovničnem označevanju korpusnih besedil (gl. prispevek Erjavec et al. 2015).[8]

```
<WordForm>
     <feat att="stopnja" val="primernik"/>
     <feat att="msd" val="Rsr"/>
     /.../
</WordForm>
```

Slika 4: Zapis pregibnih lastnosti primerniške oblike prislova *pazljivo* v formatu XML LMF.

2.4.4.3 Izrazna variantnost

Kadar enemu naboru lastnosti (eni abstraktni slovnični obliki) ustreza več oblik neke leme, govorimo o dveh ali več izraznih variantah oblik (oblikovnih dvojnicah), med katerimi ločujemo s pripisom t. i. variantnih lastnosti. V obstoječi različici leksikona so to normativne variantne lastnosti, ki označujejo zaznamovanost oblike glede na obstoječi pravopisni standard (SP 2001). Oblike brez pripisane normativne zaznamovanosti so v skladu s standardom (npr. *grad: gradu* v dajalniku ednine), medtem ko oblike s pripisom *nestandardno* niso v skladu standardom (npr. *grad: gradi* v imenovalniku množine). V primeru variantnosti dveh ali več standardnih oblik je vsem oblikam pripisana lastnost *variantno* (npr. *grad: grada* ali *grad: gradu* v rodilniku ednine).

8 Vsem primerniškim oblikam prislovov je denimo pripisana oblikoskladenjska oznaka Rsr, saj v skladu z oblikoskladenjskimi specifikacijami JOS prva črka oznake prinaša podatek o besedni vrsti oblike (R: prislov), pri čemer nato pri prislovih druga črka označuje vrsto (s: splošni), tretja pa stopnjo (r: primernik).

```
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljiveje"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="97"/>
</FormRepresentation>
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljivejše"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="2"/>
</FormRepresentation>
```

Slika 5: Zapis variantnih primerniških oblik prislova *pazljivo* z normativnimi in korpusnimi podatki v formatu XML LMF.

2.4.4.4 Korpusni podatki

Posameznemu zapisu oblike v leksikonu besednih oblik Sloleks je pripisan tudi podatek o njeni pogostosti v referenčnem korpusu. Ta je iz korpusa pridobljen avtomatsko, in sicer s poizvedbo, kolikokrat se dana oblika z dano lemo in oblikoskladenjsko oznako pojavi v korpusu. Pri presoji zanesljivosti te informacije moramo upoštevati omejitve avtomatskega označevanja korpusnih besedil, saj pojavnicam v korpusu niso nujno pripisane prava lema in slovnične lastnosti. Trenutna natančnost strojnega lematizatorja in označevalnika za slovenščino (Grčar et al. 2012), s katerim je bil označen referenčni korpus Gigafida, je 91,34 %, pri čemer se natančnost za različne skupine lem in oblik precej razlikuje (ibid.: 92–94) in lahko vpliva na ustrezno interpretacijo korpusnih podatkov (Logar et al. 2015).

2.4.5 Povezane oblike

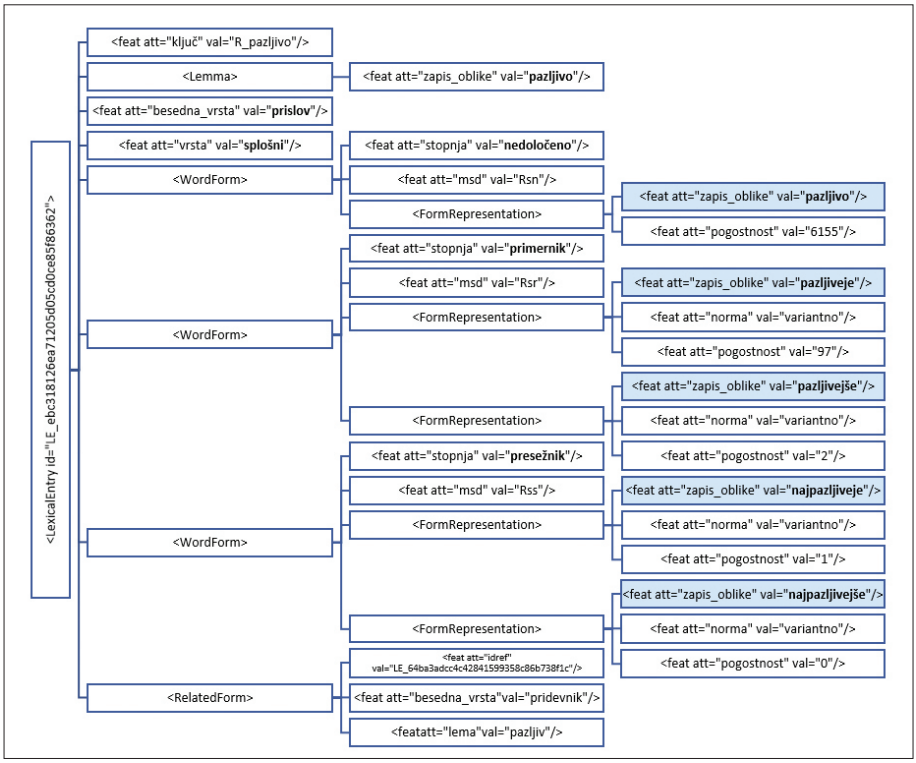
Poleg pregibnih oblikoslovnih lastnosti leme so v leksikonu besednih oblik Sloleks pri določenih skupinah lem beležene tudi informacije o njihovi besedotvorni povezanosti z drugimi lemami oz. leksikonskimi enotami. Trenutni nabor besedotvornih relacij v leksikonu besednih oblik Sloleks vključuje naslednje recipročne povezave: 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.*).

```
<RelatedForm>
     <feat att="idref" val="LE_64ba3adcc4c42841599358c8
     6b738f1c"/>
     <feat att="besedna_vrsta" val="pridevnik"/>
     <feat att="lema" val="pazljiv"/>
</RelatedForm>
```

Slika 6: Zapis povezanih iztočnic prislova *pazljivo* v formatu XML LMF.

V povzetek opisa strukture leksikonskih enot v leksikonu besednih oblik Sloleks na Sliki 7 prikazujemo celoten zapis leksikonske enote za prislov *pazljivo*, zaradi večje preglednosti tu v shematskem prikazu formata XML LMF.

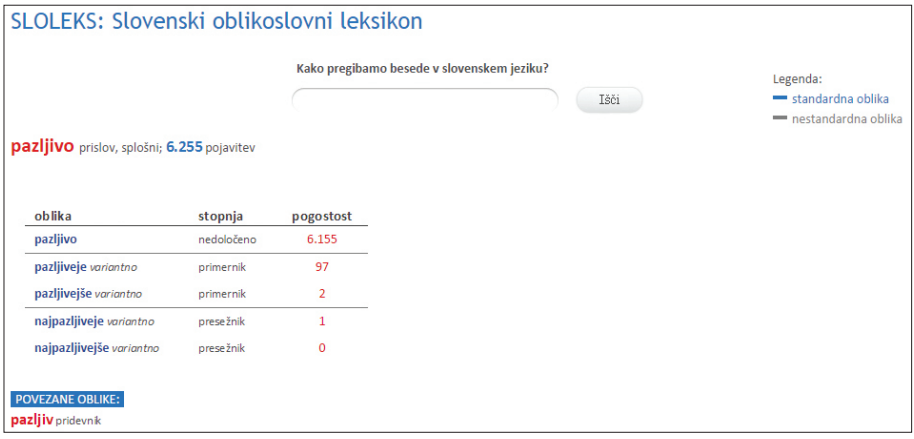


Slika 7: Prikaz strukture podatkov leksikonske enote v formatu XML LMF za prislov *pazljivo*.

2.5 Vizualizacija

Poleg uporabnosti v jezikovnotehnoloških aplikacijah strukturiran zapis informacij o oblikoslovnih lastnostih slovenskega besedišča prinaša številne prednosti tudi pri njegovi integraciji v druge priročnike ali samostojni vizualizaciji, saj nam omogoča poljubno določanje tako nabora prikazanih informacij kot tudi načina njihovega prikazovanja. Primer enega izmed načinov vizualizacije oblikoslovnega leksikona kot samostojnega oblikoslovnega priročnika je vizualizacija leksikona besednih oblik Sloleks na spletnem portalu projekta Sporazumevanje v slovenskem jeziku.[9]

Kot prikazuje posnetek prikaza leksikonske enote za prislov *pazljivo* (Slika 7) na Sliki 8, je podatek o lemi prikazan s poudarjeno pisavo, v isti vrstici pa mu sledita še podatka o besedni vrsti in vrsti ter skupna pogostost, ki v leksikonski enoti ni eksplicitno zapisana, a je izračunana na podlagi podatka o pogostosti posameznih pregibnih oblik. Sledi vizualno ločen prikaz pregibne paradigme z normativnimi in slovničnimi lastnostmi, pri čemer so posamezne kombinacije pregibnih lastnosti (slovnične oblike) ločene s črto. S klikom na podatek o pogostosti v referenčnem korpusu uporabnik dostopa do konkordanc v spletnem konkordančniku korpusa, ki jih na podlagi iskalnega pogoja v obliki dane kombinacije leme, oblike in oblikoskladenjske oznake mehanizem v korpusu poišče samodejno. Ob koncu gesla so prikazane tudi morebitne povezane leksikonske iztočnice z izpisano lemo in podatkom o besedni vrsti, ob kliku na lemo pa je uporabnik preusmerjen na ustrezno geslo.



Slika 8: Prikaz leksikonske enote za prislov *pazljivo* na spletnem portalu.

9 http://www.slovenscina.eu/sloleks (dostop 30. 6. 2015).

3 SMERNICE NADGRADNJE

3.1 Širitev geslovnika

Kot smo opisali v razdelku 2.1., leksikon besednih oblik Sloleks trenutno vsebuje približno 100.000 najpogostejših lem slovenskega besedišča in, v primerjavi z geslovniki drugih dostopnih oblikoslovnih priročnikov za slovenščino, ki so bodisi manjši po obsegu (Apertium, MULTEXT-East) ali niso nastajali na korpusni osnovi (SP 2001, SSKJ), pokriva doslej največji delež splošnega besedišča slovenskega jezika. Toda kljub temu je ob načrtovanju slovarskih in drugih jezikoslovnih opisov sodobne slovenščine na eni strani in upoštevanju naraščajočih in raznolikih potreb njenega strojnega procesiranja na drugi nujna tudi njegova nadaljnja širitev. To je smiselno načrtovati v obliki treh koncentričnih krogov, pri čemer vsak izmed krogov predstavlja temeljno izhodišče naslednjega, ni pa nujno, da so pri njihovi implementaciji uporabljena enaka vsebinska ali metodološka izhodišča.

V kontekstu prioritetne integracije oblikoslovnih podatkov v digitalno zasnovane priročniške opise sodobne slovenščine prvi koncentrični krog nadaljnjih širitev leksikona besednih oblik Sloleks predstavlja njegovo usklajevanje z geslovnikom slovarske baze, torej vključitev (manjkajočih) jedrnih leksikalnih enot slovenskega jezika, vključno z večbesednimi slovarskimi iztočnicami, variantnimi zapisi in drugimi lemami oz. oblikami, ki so oblikoslovno povezane z lemo neke slovarske iztočnice.

Drugi krog širitve geslovnika leksikona besednih oblik vključuje besedišče referenčnega korpusa slovenskega jezika. To glede na merila za zajem iztočnic namreč ni nujno tudi del slovarskega geslovnika, a je zaradi svoje pogostosti v rabi ključnega pomena za razvoj jezikovnih tehnologij, tudi tistih, ki se uporabljajo pri izdelavi slovarja, saj morajo lematizatorji, slovnični označevalniki in luščilniki leksikalnih podatkov poleg slovarske iztočnice natančno prepoznavati tudi besedišče v njeni okolici. V skladu s krepostnim krogom jezikoslovnega označevanja se s širitvijo leksikona izboljša jezikovni model označevalnika, z njim pa natančnost označevanja korpusa.

Primerjava prekrivnosti besednih oblik (različnic)[10] leksikona besednih oblik Sloleks in besedišča referenčnega korpusa Gigafida razkriva, da Sloleks vsebuje zgolj 43 % različnic, ki se v korpusu Gigafida pojavijo vsaj petkrat. Z višanjem frekvenčnega praga ta delež pričakovano narašča, vendarle pa leksikon besednih oblik pokriva zgolj 79 % od skupno 251.292 različnic, ki se v korpusu pojavijo najmanj 100-krat. Taka pogostost posamezne različnice (torej oblike, ne leme) v uravnoteženem in reprezentativnem korpusu jezika narekuje tudi potrebo po njenem formalnem opisu v ustrezni podatkovni bazi.

10 Pri tem smo namenoma primerjali zgolj besedne oblike z zapisom z malimi črkami, saj se nismo želeli opirati na strojno pripisan podatek o lemi ali posebnosti zapisovanja v korpusnih besedilih (npr. *slovenija*, *ljubljana* ipd.).

Podrobnejša analiza seznama najpogostejših različnic v korpusu Gigafida, ki jih v leksikonu besednih oblik Sloleks še ni, kaže, da bi to bazo v prihodnje veljalo dopolniti predvsem z naslednjimi skupinami besedišča:

- različnimi krajšavami (*p., s., j..; nan., dok., mr.; m2, cm3, a3; UV, MMS, VIP, SUV; VPS, SŽ* itd.),
- prevzetimi samostalniki (*city, miss, fax, art, dj, bluetooth, mac, facebook, prix, alias, maestro, college, gay, styling, fitness, volley, weekend, hiphop* itd.),
- nesklonljivimi prilastki (*turbo, online, anti, stereo, retro, audio, etno, latino, afro* itd.),
- nestandardnimi oblikami (*tud, kr, blo, brezveze, dobr, nevem, kao, jst, jap, tolk, nč, lahk, drgač, al, tm, zarad, mislm, pomoje, una, brezveze* itd.),
- medmeti (*živjo, bognedaj, jao, jp, hehe, he, hahaha, hahahaha, sviš, hehehe, khm* itd.),
- tujimi in domačimi lastnimi imeni (*obama, ilirika, evroliga, barca, clio, patria, beverly, pomurec, messi, airways, michel, svena, sarkozy, coca, evrovizija, titanik, čedad, wikipedia* itd.),
- zvrstno zaznamovanim besediščem (*škrinja, zalüblenih, mojoga, škürec, zadvečerek, špas* itd.),
- pa tudi z nekaterimi pogostimi domačimi oz. povsem podomačenimi besedami (*drugouvrščen, mimoidoči, prida, kapitalov, superpokal, štoparski, fotogalerija, tričetrt, bogve, drugoligaški, didžej, avtohiša, enoprostorec, osemvaljnik, supermodel, drska, preska, četrtinski, požarnik, klaviaturist, klientelizem, kapetanski, avtoprevozništvo, označba, predizbor, napak, prismučati, nezemljan, brezplačnik, evroobmočje, streljaj, dvetretjinski* itd.).

Za potrebe jezikovnih tehnologij je smiselno načrtovati tudi popis tujega besedišča, ki v slovenščino ni bilo prevzeto, a se v slovenskih besedilih pogosto pojavlja, denimo kot del tujih stvarnih lastnih imen (npr. *the, of, and* itd.).

Po geslovniku slovarske baze in pogostem besedišču referenčnega korpusa tretji krog razširitve geslovnika predvideva vključitev specializiranega besedišča za potrebe specifičnih jezikovnih priročnikov ali tehnoloških aplikacij, denimo tipično govorjenega besedišča, besedišča posameznih strokovnih področij, narečnega besedišča ali drugega zvrstno zaznamovanega besedišča. Za razliko od prvih dveh krogov, ki predstavljata univerzalno jedro opisa leksike nekega jezika, te tretje, domensko pogojene, širitve leksikona ni mogoče predvideti ali zagotoviti vnaprej. Ključnega pomena pa je, da je skupnosti omogočeno samostojno dopolnjevanje jedrnega leksikona, vključno z orodji in viri, ki jih za to potrebuje, začenši z odprtodostopno bazo pregibnih vzorcev.

3.2 Formalizacija oblikoslovnih vzorcev

Eno najpomembnejših vprašanj, povezanih tako s širitvijo kot reevalvacijo obstoječih oblikoslovnih leksikonov za slovenščino, je izdelava nabora strojno berljivih vzorcev pregibanja besed v slovenskem jeziku, ki bi omogočil validacijo pregibnih paradigem iztočnic v obstoječih priročnikih, pripisovanje paradigem novim lemam ter razvoj metod za njihovo samodejno prepoznavanje v besedilnih korpusih (prim. npr. Šnajder 2013 za hrvaščino).

Ob naboru oblikoslovnih leksikonov za slovenski jezik je mogoče sklepati, da v slovenskem prostoru že obstaja več tovrstnih zbirk pregibnih vzorcev, toda te širši raziskovalni skupnosti niso dostopne, načela njihovega oblikovanja, razvrščanja in usklajevanja z jezikovno rabo pa večinoma niso dokumentirana. Poskus implicitnega opisa vzorcev na podlagi luščenja pregibnostnih paradigem v večjih dostopnih podatkovnih zbirkah ali priročnikih, kot so SP2001, Apertium, pa tudi Sloleks (K. Dobrovoljc 2014), po drugi strani razkriva tudi nesistematičnosti in gradivne pomanjkljivosti, saj se v vseh virih pri naboru in razvrščanju vzorcev pojavljajo napake, nedoslednosti ali neskladnosti s sodobno jezikovno rabo.

To potrjuje, da je vsakršno načrtovanje izrabe ali nadgradnje obstoječih oblikoslovnih podatkovnih zbirk neločljivo povezano tudi z izdelavo aktualiziranega, prosto dostopnega seznama formaliziranih vzorcev pregibanja v slovenskem jeziku. Za razliko od tradicionalnih jezikoslovnih pristopov k opisu vzorcev pregibanja v slovenščini pa njihova jezikovnotehnološka namembnost narekuje upoštevanje spodaj opisanih načel ločene obravnave oblikovnih in izgovornih vzorcev, strojne berljivosti ter usklajenosti z jezikovno rabo.

3.2.1 Ločevanje oblikovnih in naglasnih vzorcev

Referenčni jezikovni priročniki za slovenščino tradicionalno pregibne izrazne lastnosti slovenskega besedišča opisujejo s hkratnim opazovanjem oblikovnih in izgovornih sprememb pri pregibanju, z razvrščanjem v t. i. sheme za dinamični naglas in oblikoslovje oz. sheme za tonemski naglas (prim. Rigler v SSKJ: XXXVIII–XLIX, LV–LVIII). Čeprav je tovrstno privzeto prikazovanje onaglašene paradigme delno vprašljivo tudi z vidika informativnosti za uporabnike (zlasti tuje, ki onaglaševanja niso vajeni), je predvsem z vidika uporabnosti v jezikovnotehnoloških aplikacijah smiselno pri opisu pregibnih lastnosti slovenskega jezika vzpostaviti ločnico med oblikoslovnimi vzorci v pisnem in naglasnimi vzorci v govorjenem jeziku. Z vidika strojnega procesiranja naravnih jezikov gre namreč

za dve ločeni ravni opisa (in procesiranja), ki sta medsebojno povezani, a izrazno neodvisni, saj se v jezikovni rabi vedno materializira le ena izmed obeh izraznih možnosti (zapis ali izgovor).

3.2.2 Strojna berljivost vzorcev

Drugi ključni vidik formalizacije pregibnih vzorcev je strojna berljivost njihovega zapisa, ki za razliko od diahrono ali pomensko motiviranih jezikoslovnih opisov vzorcev pregibanja v slovenščini zahteva skrajno formalistično pojmovanje osnove kot tistega nespremenljivega niza grafemov ali fonemov, ki je skupen vsem besednim oblikam v paradigmi, in obrazila kot tistega niza, ki je tej osnovi (krnu) dodan za tvorbo posamične oblike, ne glede, ali je dodan na koncu, začetku ali celo znotraj osnove. Strojno berljivi vzorec je tako zapisan v obliki algoritmičnih pravil, ki določajo načine krnjenja leme v osnovo in tvorjenja pregibnih oblik iz te osnove. Primer opisa pregibnih vzorcev za pregibanje po prvi ženski sklanjatvi, ki je prevedljiv v poljubni programski jezik, prikazuje Tabela 2.

To pomeni, da so tudi modifikacije ali posebnosti splošnih vzorcev, kot so premene, enoštevilske paradigme, raznospolske sklanjatve ipd., obravnavane kot samostojni pregibni vzorci. Z vidika njihove distribucije pa je nato smiselno ločevati med produktivnimi (sistemskimi) vzorci (ki se lahko povezujejo z odprto množico lem) in vzorci za izjeme (ki se povezujejo z zelo omejenim naborom lem). Tako produktivni vzorci kot vzorci za izjeme so formalizirani na enak način, informacija o njihovi omejeni distribuciji pa je lahko vzorcu pripisana v obliki neobvezujoče dodatne informacije ali v obliki restrikcije.[11]

Tabela 2: Primer opisa treh pregibnih vzorcev za samostalnike ženskega spola.

	vzorec ***lipa***		vzorec ***mravlja***		vzorec ***gora***	
osnova	lemi odstrani zadnjo črko	*lip*	lemi odstrani zadnjo črko	*mravlj*	lemi odstrani zadnjo črko	*gor*
Sozei	osnova + a	*lipa*	osnova + a	*mravlja*	osnova + a	*gora*
Sozer	osnova + e	*lipe*	osnova + e	*mravlje*	osnova + e	*gore*
Sozed	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*
Sozet	osnova + o	*lipo*	osnova + o	*mravljo*	osnova + o	*goro*
Sozem	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*

11 Na enak način lahko za potrebe različnih raziskav ali aplikacij vključimo tudi druge tipe metapodatkov, npr. o pričakovanih izraznih in slovničnih lastnostih lem, ki se pregibajo po določenem vzorcu, o medsebojni povezanosti vzorcev, o povezavah z naglasnimi vzorci ipd.

	vzorec ***lipa***		vzorec ***mravlja***		vzorec ***gora***	
Sozeo	osnova + o	*lipo*	osnova + o	*mravljo*	osnova + o	*goro*
Sozdi	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*
Sozdr	osnova	*lip*	vrini *e* pred zadnji dve črki osnove	*mravelj*	osnova osnova + a	*gor* *gora*
Sozdd	osnova + ama	*lipama*	osnova + ama	*mravljama*	osnova + ama	*gorama*
Sozdt	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*
Sozdm	osnova + ah	*lipah*	osnova + ah	*mravljah*	osnova + ah	*gorah*
Sozdo	osnova + ama	*lipama*	osnova + ama	*mravljama*	osnova + ama	*gorama*
Sozmi	osnova + e	*lipe*	osnova + e	*mravlje*	osnova + e	*gore*
Sozmr	osnova	*lip*	vrini *e* pred zadnji dve črki osnove	*mravelj*	osnova osnova + a	*gor* *gora*
Sozmd	osnova + am	*lipam*	osnova + am	*mravljam*	osnova + am	*goram*
Sozmt	osnova + e	*lipe*	osnova + e	*mravlje*	osnova + e	*gore*
Sozmm	osnova + ah	*lipah*	osnova + ah	*mravljah*	osnova + ah	*gorah*
Sozmo	osnova + ami	*lipami*	osnova + ami	*mravljami*	osnova + ami	*gorami*

Za ponazoritev razmerja med jezikoslovno in strojno motiviranimi opisi oblikoslovnih vzorcev v slovenščini vzemimo primer neonaglašenega pregibanja ženskih samostalnikov po prvi ženski sklanjatvi. Po Riglerjevih oblikoslovno-naglasnih shemah tej sklanjatvi ustreza pet shem (IAb1, IAb3, IB1b1, IB2a1 in IB2b1), pri čemer ima vsaka izmed shem enega ali več podtipov, oblikam v posameznih podtipih pa so lahko pripisane nadaljnje opombe z opisi ene ali več možnih modifikacij. Preslikavo razvezave tovrstnih oblikovno-naglasnih vzorcev v formalne vzorce pregibanja prikazujemo v Tabeli 3.[12] Vidimo lahko, da razmerje med obema načinoma opisa ni simetrično, saj imajo lahko posebnosti posameznih tipov drugačen formalni opis kot izhodiščni tip (npr. *ladja* drugače kot *lipa*, ali *mravlja* drugače kot *tabla*), prav tako pa ob ločevanju na oblikovne in naglasne vzorec nekateri naglasno razlikovalni tipi združijo v skupni oblikovni vzorec (npr. *gôra* in *stezà* v *steza*).

12 V primerjavo nismo vključili pridevniškega pregibanja (IAb1-podtip *désne*), tipa IAb1-agape in vzorcev za pregibanje po enem samem številu (*vile, grablje, nečke* ipd.).

Tabela 3: Primer preslikave Riglerjevih shem v formalizirane vzorce pregibanja po prvi ženski sklanjatvi.

Riglerjeve sheme		Formalni vzorci
1	shema IAb1, podtip *lípa*	vzorec ***lipa***
2	shema IAb1, podtip *lípa*, opomba 8, primer *óboj*	vzorec ***oboa***
3	shema IAb1, podtip *lípa*, opomba 8, primer *dékel*	vzorec ***dekla***
4	shema IAb1, podtip *lípa*, opomba 8, primer *grábelj*	vzorec ***mravlja***
5	shema IAb1, podtip *lípa*, opomba 8, primer *ládij*	vzorec ***ladja***
6	shema IAb1, podtip *lípa*, opomba 8, primer *kámer*	(vzorec *dekla*)
7	shema IAb1, podtip *lípa*, opomba 8, primer *zárij*	(vzorec *ladja*)
8	shema IAb1, podtip *lípa*, opomba 8, primer *mrávelj/márenj*	(vzorec *mravlja*)
9	shema IAb3, podtip *búkev*	vzorec ***bukev***
10	shema IAb3, podtip *cerkvé*	vzorec ***cerkev***
11	shema IB1b1, podtip *stezà*	vzorec ***steza***
12	shema IB1b1, podtip *stezà,* opomba 19	(vzorec *steza*)
13	shema IB1b1, podtip *stezà,* opomba 20	(vzorec *steza*)
14	shema IB2a1, podtip *gospá*	vzorec ***gospa***
15	shema IB2b1, podtip *gôra*	(vzorec *steza*)
16	shema IB2b1, podtip *nečké*	(vzorec *lipa*)

Kot smo izpostavili, bi morali za potrebe formalizacije na enak način opisati tudi vzorce za izjeme, ki se povezujejo zgolj s posameznimi lemami (npr. *ovca, mati, hči,* ipd.) in v obstoječih priročnikih niso omenjene kot del shem, temveč v geselskem zaglavju (npr. *ôvca -e stil. -é ž, rod. mn. ôvc in ovác (ó)* v SSKJ2). Njihov zapis v obliki samostojnega vzorca je poleg sistematičnosti smiseln tudi zato, ker tudi vzorci za izjeme izkazujejo določeno mero sistemskosti in se običajno povezujejo z več kot zgolj eno samo lemo (npr. *deska* tako kot *ovca*, *pramati* tako kot *mati* ipd.).

3.2.3 Skladnost z jezikovno rabo

Tretje ključno načelo, ki ga je treba upoštevati pri izdelavi nabora formaliziranih oblikoslovnih vzorcev za slovenščino, pa je upoštevanje oblikoslovnih tendenc v sodobni jezikovni rabi, o kakršni lahko sklepamo na podlagi uravnoteženih in reprezentativnih korpusov sodobne slovenščine. Z vidika jezikovnih tehnologij je upoštevanje jezikovne rabe pomembno zaradi zagotavljanja čim večje pokritosti pogosto rabljenega besedišča (ne glede na njegovo skladnost z obstoječo

kodifikacijsko normo), za jezikoslovje pa korpusni pristop k opisu oblikoslovja slovenskega jezika ponuja priložnost za aktualizacijo obstoječih opisov, ki niso nastali na tako obsežni gradivni osnovi.

Analiza rabe lahko prinaša nova spoznanja že na ravni opisa vzorcev pregibanja. Če za primer vzamemo zgolj zgoraj opisani nabor vzorcev za prvo žensko sklanjatev, analiza rabe v korpusu Gigafida denimo pokaže, da se domnevno sistemski podtip sheme IAb3, po katerem naj bi se samostalniki ženskega spola na *-ev* v rodilniku dvojine ali množine pojavljali tudi s končnico -á (*cérkev*: *cérkev* in *cerkvá*) pojavlja zgolj pri ponazoritveni lemi (gre torej za izjemo in ne pravilo), prav tako pa se v tožilniku dvojine v korpusu ne pojavlja navedena stilistična oblika *cerkvé*. Podobno analiza korpusa pod vprašaj postavlja tudi trditev, da se v rodilniku dvojine in množine med dva zvočnika *e* vriva samo v primeru, kadar je drugi zvočnik *r* (*kamra: kamer*), saj raba kaže, da do vrivanja *e* lahko prihaja tudi pri nekaterih drugih zvočniških sklopih (npr. *himna: himen*, *kolumna: kolumen*; *avla: avel*).

Še zlasti pa je analiza rabe pomembna z vidika oblikoslovnega razvrščanja leksike, torej povezovanja konkretnih lem s konkretnimi vzorci pregibanja. Za primer vzemimo obrazilno stopnjevanje prislovov, kjer slovenščina pozna dva sistemska vzorca: bodisi se prislovi ne stopnjujejo bodisi se stopnjujejo z variantnima končnicama *-ejše* oz. *-eje*, če so tvorjeni iz pridevnika in njihov pomen to dopušča (*zanimivo: zanimivejše/zanimiveje*). Analiza razvrščanja obeh sistemskih vzorcev v leksikonu besednih oblik Sloleks in slovarju Slovenskega pravopisa (K. Dobrovoljc 2014) kaže, da v obeh priročnikih prihaja do razhajanj z rabo, saj so nekateri v rabi stopnjevani prislovi v enem ali obeh priročnikih navedeni brez stopnjevanih oblik (npr. *smiselno, preudarno, poredko, enakovredno, korektno, športno* itd.) ali pa je stopnjevanost pripisana tistim prislovom, ki v rabi ne izkazujejo nobenega izmed variantnih obrazil (*arogantno, bistroumno, strahovito, zagonetno* itd.).

Še posebej pa mora biti skladnost s sodobno jezikovno rabo merilo pri določanju izjem. Pri stopnjevanju prislovov tako Slovenski pravopis za prislove *drago, ozko* in *težko* denimo navaja oblike *dražje, ožje* oz. *težje* (čeprav so v rabi tudi oblike *draže, ože* oz. *teže*), za prislov *kratko* obliki *krajše* in *kračje* (čeprav slednje v korpusu ni), za prislov *gladko* oblike *gladkeje, gladkejše, glaje* in *glajše* (čeprav se oblika *glaje* v referenčnem korpusu ne pojavi, oblika *glajše* pa samo enkrat) in tako dalje.

3.3 Dodajanje izgovora

Leksikon besednih oblik Sloleks trenutno ne vsebuje podatkov o izgovornih lastnostih vsebovanih besednih oblik, kar pomeni, da so zapisi osnovnih in

pregibnih oblik neonaglašeni. Z namenom celovitega opisa izraznih lastnosti sodobnega slovenskega besedišča je torej vključitev podatkov o izgovoru besednih oblik ena izmed prioritetnih nadgradenj obstoječe različice leksikona besednih oblik Sloleks. To je še toliko bolj pomembno z vidika govornih tehnologij, saj v slovenskem prostoru trenutno ne obstaja prosto dostopni leksikon, ki bi omogočal razvoj strojnih razpoznavalnikov in sintetizatorjev govora za potrebe najrazličnejših aplikacij, kot so samodejni podnaslavljalniki, bralniki za slepe in slabovidne, sistemi za komuniciranje v naravnem jeziku ipd.

3.3.1 Fonetični zapis

Podatek o izgovoru bi moral biti v leksikonu besednih oblik Sloleks beležen v obliki fonetične transkripcije v dogovorno izbrani strojno berljivi mednarodni fonetični abecedi (prim. Jurgec 2015), ki poleg zapisa glasov inherentno vsebuje tudi podatek o mestu in jakosti naglasa. Glede na specifične jezikovnopriročniške ali jezikovnotehnološke potrebe se referenčni fonetični zapis lahko na podlagi pravil pretvori tudi v druge fonetične abecede ali zapise onaglašenih oblik. Ti so lahko, ni pa nujno, v obliki samostojnih elementov zabeleženi tudi v samem leksikonu besednih oblik, pomembno pa je, da so ves čas usklajeni z izgovorom v referenčni izhodiščni fonetični transkripciji. Na enak način je mogoče vključevati tudi podatek o tonemskem naglasu, ki sicer glede na potrebe strojnih in človeških uporabnikov ni prioriteten (gl. Arhar et al. 2015).

3.3.2 Strukturna umestitev

Kot smo že opisali v razdelku 3.2.3, v leksikonu besednih oblik Sloleks izgovor obravnavamo kot osamosvojeno informacijo o glasovni podobi posamezne neonaglašene pisne oblike. Podatek o izgovoru tako pripisujemo na ravni obstoječega zapisa osnovne oz. pregibnih oblik. V primeru v slovenščini zelo pogoste izgovorne variantnosti je lahko eni neonaglašeni obliki pripisanih tudi več elementov s podatkom o izgovoru, med njimi pa tako kot pri variantnosti neonaglašenih oblik ločujemo z ustreznimi kvalifikatorji, ki nam omogočajo samodejni priklic izgovora posamezne ali vseh oblik v eni izmed variantnih izgovornih paradigem (glej razdelek 3.4). Na ta način obravnavamo vse tipe izgovorne variantnosti, ne glede na to, ali gre za glasovno (prevajalka: *prevajalka-pravajaụ̯ka*) oz. naglasno variantnost (agencija: *agéncija-agencíja*) vseh ali zgolj ene izmed pregibnih oblik v paradigmi.

3.3.3 Obravnava enakopisnic

Tudi po vključitvi informacij o izgovoru so leme z enakim zapisom in izgovorom ločene v več samostojnih leksikonskih enot, če izkazujejo različne izrazne lastnosti, tj. spadajo v različne besedne vrste (prislov in pridevnik *spet*), imajo različne leksikalne lastnosti (ženski in moški samostalnik *prst*) ali se drugače pregibajo (*vesti*: *vedem* in *vesti: vezem*). Po drugi strani v leksikonu besednih oblik Sloleks ne ločujemo med homonimnimi lemami s povsem prekrivnimi izraznimi, a različnimi pomenskimi lastnostmi (npr. moški samostalnik *bor*-drevo ali *bor*-element), zato take pare leksemov še naprej obravnavamo kot eno samo izrazno enoto besedišča, ki ji ustreza ena sama leksikonska enota (samostalnik moškega spola *bor*).[13] Ob dejstvu, da v leksikonu besednih oblik ne beležimo tonemskosti, enako velja tudi za pare pomensko različnih enakopisnic, ki se razlikujejo zgolj v tonemskem naglasu (pregibanje pridevnikov *bȗčen*-nanašajoč se na bučo in *búčen*-glasen opišemo v skupni leksikonski enoti splošnega pridevnika *bučen*).[14]

Po drugi strani pa vključevanje informacij o izgovoru spreminja način obravnave lem z enakim zapisom in drugačnim izgovorom, ki so bili doslej v primeru prekrivne leme, slovničnih lastnosti in neonaglašene pregibne paradigme obravnavani kot ena sama leksikonska enota, npr. *partija* (*partíja* in *pártija*), *častiti* (*částiti* in *častíti*) itd. Z uvedbo pomensko razločevalnih podatkov o izgovoru se oba leksema osamosvojita v samostojni leksikonski enoti (*S_partija_1* in *S_partija_2*), ob tem pa moramo upoštevati, da trenutni slovnični označevalniki za slovenščino ne vključujejo pomenskega razdvoumljanja oblikovno prekrivnih enakopisnic v kontekstu, zato so pojavnice tovrstnih parov označene z identično lemo in oblikoskladenjsko oznako. To pomeni, da je v leksikonu besednih oblik enakim oblikam ene in druge leme pripisana enaka korpusna pogostnost, pri luščenju informacij o besedilnem okolju, zgledov in drugih vrstah besedilnih podatkov pa med njima ni mogoče ločevati zgolj z avtomatskimi postopki.

3.4 Kategorizacija variantnosti

Oblikoslovna variantnost, tj. več izraznih možnosti iste slovnične oblike, je v slovenščini zelo pogosta in do nje prihaja na različnih ravneh: pri zapisu (*v naprej* ali *vnaprej*), izgovoru glasov (*/drsáuka/* ali */drsálka/*) ali naglaševanju leme (*upokójenec* ali *upokojênec*), kakor tudi pri izbiri oblikoslovne paradigme (*Luka: Luka, Luke* ali *Lukata*), zapisa ali izgovora pregibnih oblik (*college*: *collegea* ali *collega*) ter besedotvorju (*vanilija: vanilijev, vanilijin* ali *vanilin*).

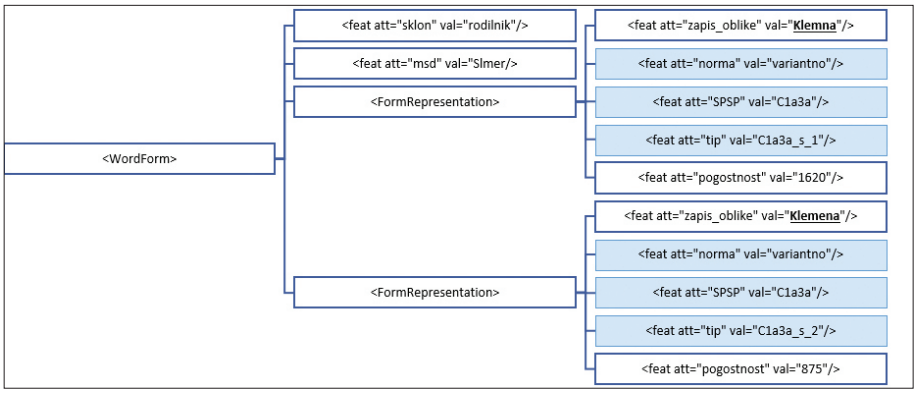
13 Za razmerje med leksikonsko in slovarsko iztočnico gl. prispevek K. Dobrovoljc (2015).

14 Za razumevanje preslikav med formalnooblikoslovno motiviranimi leksikonskimi enotami in pomensko motiviranimi slovarskimi enotami gl. Gantar (2015) in K. Dobrovoljc (2015).

Če bi v rabi izkazane oblikoslovne variante v leksikon besednih oblik vključevali zgolj kot niz več zapisovalnih ali izgovornih oblik z identičnimi slovničnimi lastnostmi, med njimi brez dodatnih pravil ne bi mogli sistematično ločevati, zato je z vidika različnih namenov uporabe oblikoslovnih leksikonov koristno, da variantam pripišemo tudi njihove razlikovalne (variantne) lastnosti oz. jih ustrezno tipologiziramo. Pri tem je treba poudariti, da tipologizacije ne smemo enačiti z normativno kvalifikacijo, saj prva označuje jezikovnosistemsko izbiro, druga pa njeno naknadno jezikoslovno interpretacijo, ki je družbenokonsenzualno pogojena in s tem bolj ali manj spremenljiva. Obe informaciji sta v leksikonu besednih oblik nepogrešljivi, saj tipologizacija omogoča usmerjen priklic posamičnih variantnih oblik ali celotnih variantnih paradigem ene ali več leksikonskih enot, podatek o njihovi normativni (ne)zaznamovanosti pa je ključen pri integraciji leksikona v jezikovne priročnike, navsezadnje pa informacijo o (ne)standardnosti posameznih variantnih izbir potrebujejo tudi jezikovnotehnološke aplikacije, ki generirajo besedila v slovenščini, kot so strojni prevajalniki, sintetizatorji govora ipd.

Dodajanje podatkov o tipih in normativni zaznamovanosti variantnih oblik smo v leksikonu besednih oblik Sloleks že preizkusili pri vzpostavitvi zasnove in delotoka spletnega portala Slogovni priročnik, namenjenega reševanju najpogostejših jezikovnih zadreg pri tvorbi besedil v slovenščini s sopostavljanjem informacije o veljavnem pravopisnem standardu in korpusnih podatkov (Krek 2012; Krek et al. 2013a; K. Dobrovoljc in Krek 2013). Zaledni mehanizem, ki uporabnikovo vprašanje poveže z ustrezno jezikovno zadrego in njenim pojasnilom ter vizualizira korpusne in normativne podatke za konkretno obliko ali paradigmo, po kateri sprašuje uporabnik, vse potrebne podatke črpa iz leksikona besednih oblik Sloleks, v katerem so leme ali pregibne oblike, povezane z eno izmed obravnavanih jezikovnih zadreg, ustrezno kategorizirane. Vsaki obliki (osnovni ali pregibni) so tako pripisani trije tipi kategorizacijskih podatkov: (i) kategorija jezikovne zadrege oz. variantne izbire, ki temelji na ontološko urejenem seznamu jezikovnih zadreg v slovenščini (H. Dobrovoljc in Krek 2011; Bizjak Končar et al. 2011), (ii) tip sistemske variante znotraj kategorije in (iii) njena normativna vrednost.

Primer take kategorizacije prikazuje izsek zapisa leksikonske enote za samostalnik *Klemen* na Sliki 9. Leksikonski enoti je že na prvem nivoju pripisan podatek, da se povezuje z jezikovno zadrego C1a3a (*Oblikoslovje* > *Samostalniki* > *Moške sklanjatve* > *Samostalniki z neobstojnim samoglasnikom* > *Slovenska lastna imena*), posameznim oblikam v paradigmi pa je poleg podatka, da spadajo v to kategorijo variantnosti v elementu z opredelitvijo tipa pripisan še podatek, kateri izmed obeh variant pripada (C1a3a_s_1 denimo označuje paradigmo z izpustom *e*, C1a3a_s_2 pa paradigmo brez izpusta), v elementu z opredelitvijo norme pa še ustrezni normativni kvalifikator *variantno,* ki označuje standardno dvojnico.



Slika 9: Del leksikonske enote *S_Klemen* s pripisom kategorije variantnega sklanjanja.

Tovrstna kategorizacija variantnosti v bazi nam tako po eni strani omogoča nadzorovan priklic podatkov posamezne leksikonske enote, kot je denimo seznam vseh jezikovnih zadreg, s katerimi se ta povezuje, ali ene oz. več oblik njene variantne paradigme, po drugi strani pa nam omogoča tudi samodejni priklic seznama vseh drugih lem, ki izkazujejo enako vrsto oblikovne, izgovorne ali besedotvorne variantnosti, npr. vseh slovenskih lastnih imen z neobstojnim samoglasnikom.

4 ZAKLJUČEK

Leksikon besednih oblik Sloleks z oblikoslovnimi, besedotvornimi, izgovornimi, normativnimi in drugimi podatki predstavlja vezni člen med različnimi jezikovnimi viri, ki jih predvideva Predlog za izdelavo slovarja sodobnega slovenskega jezika (Krek et al. 2013b). Predvsem so to različni jezikoslovno označeni korpusi slovenščine, od referenčnega, uravnoteženega, govornega do zgodovinskega in vseh ostalih. Na drugi strani so podatki iz leksikona neposredno uporabni in uporabljeni v priročniških virih, kot je (digitalni) slovar, spletni slogovni priročnik in drugi. Z enotno obravnavo morfologije slovenskega jezika leksikon v celotni označevalni in priročniški ekosistem prinaša konsistentno obravnavo pojavov, kar sicer v tem trenutku predstavlja enega ključnih primanjkljajev tako znotraj računalniškega procesiranja slovenščine kot tudi poučevanja slovenščine v celotnem izobraževalnem procesu, od osnovnih in srednjih šol do poučevanja slovenščine kot tujega jezika.

Obstoječi leksikon besednih oblik Sloleks predstavlja dobro osnovo za nadgradnjo, ki obsega predvsem obsežno širjenje geslovnika in oblik, vključno z

večbesednimi enotami, ter dodajanje informacij o izgovoru in normi. Vsi ti procesi morajo biti čim bolj avtomatizirani, kar obstoječe tehnologije že zagotavljajo, manjka pa strojno berljiv in na sodobni jezikovni rabi utemeljen popis pregibnih vzorcev za slovenščino. Nadgrajevanje je treba razumeti kot permanenten proces brez končne točke v času, saj v jezik vedno prihajajo nove besede, ki jih je treba tako opisati kot tudi zagotoviti njihovo strojno obdelavo v avtentičnih besedilih. Na ta način je koncept leksikona besednih oblik Sloleks tudi zastavljen. Najbolj ključen pri zagotavljanju širše rabe leksikona je dostop pod odprto licenco, saj šele ta zagotavlja, da se investicija v njegov razvoj tudi upraviči, v širši perspektivi pa slovenščini zagotavlja možnost za obstanek tudi v prihajajočem, vse bolj digitaliziranem svetu.

Normativna informacija v sodobnem slovarju

Damjan Popič

Abstract
This paper presents the concept of an online Slovenian Style Guide – an alternative language guide for Slovenian, the main focus of which is to equip language users for independent and full participation in the modern information society. This paper presents the structure and content of the guide, as well as its place and role within the Slovenian cultural environment. The modern guidelines for compiling language guides which are set out by the requirements of the modern information society are compared to the historic features of Slovenian linguistic prescription. In this paper, details are presented of the aim, content and construction of the Slovenian Style Guide, and the relationship between this guide and the concept of a new, dictionary of contemporary Slovenian is established.

Keywords: style guide, standardisation, orthography, language guides, language technology

Ključne besede: slogovni priročnik, standardizacija, pravopisje, jezikovni priročnik, jezikovne tehnologije

1 UVOD

V prispevku predstavimo konkretne rešitve za podajanje normativnih informacij v slovarju sodobnega slovenskega jezika v digitalnem okolju.[1] Predlog za izdelavo Slovarja sodobnega slovenskega jezika (Krek et al. 2013b) in normativnostni koncept za novi slovar (Gorjanc et al. 2015), razumeta normativnost nekoliko drugače, kot je tradicija pri nekdanjih leksikografskih izdelkih za slovenski jezik oziroma v slovenski normativni tradiciji nasploh:

> Odločitev za ločevanje normativnih od izhodiščno leksikalnih podatkov, kamor sodi zlasti pomenska členitev in prikaz tipičnega okolja besede v obliki kolokacij, je povezana s tem, da so denimo **posamezne oblike besede lahko tudi glede na obstoječi standard nesprejemljive, hkrati pa v slovenskem besedišču opazno prisotne in zato zastopane tudi v slovarju**.
>
> **Pristop, ki nestandardnih oblik ne izključuje iz obravnave, uporabniku zagotavlja, da nestandardnost sama na sebi ne pomeni hkrati tudi odsotnosti informacije** same oz. še več, uporabniku pokaže, kaj in zakaj je v določenih okoliščinah bolj ali manj tipično in/ali bolj ali manj sprejemljivo (Krek et al. 2013b: 42; poudaril D. P.).

To pomeni, da je domet normativnosti bistveno širši kot pri golem odločanju o pravilnosti in nepravilnosti posameznih variant, kot je veljalo v pretekli slovenski kodifikaciji, katere stanje lahko strnemo s citatom Jožeta Toporišiča (1991: 401), ki kljub starejšemu datumu v veliki meri velja še danes:

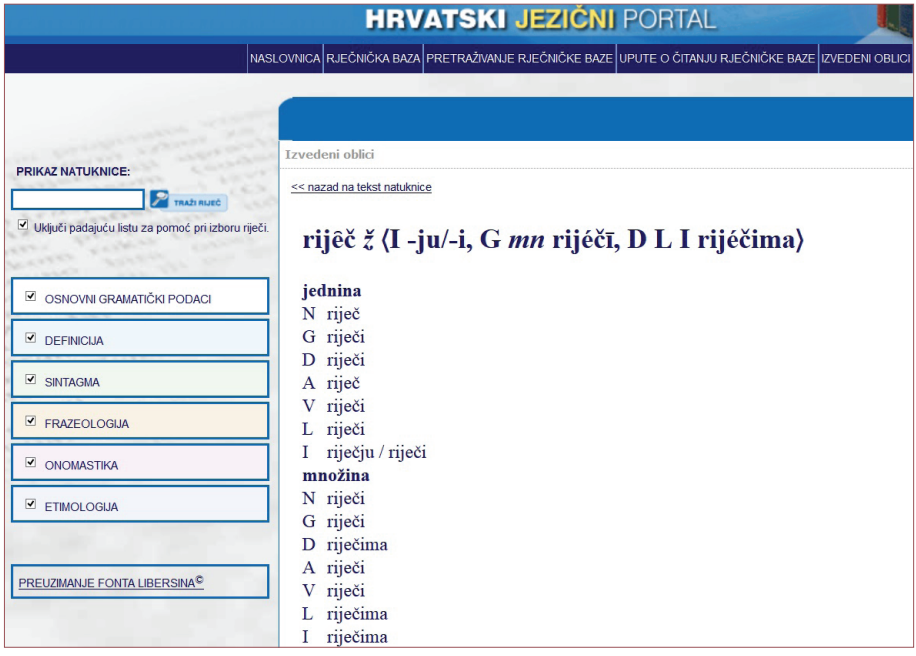
> Jasna kodifikacija se opravlja z jezikovnimi priročniki, kot so pravopis, slovnica, slovar. Pri tem je za Slovence postalo nekako značilno, da jim je predpis predvsem pravopis, ki pa je hkrati tudi pravorečje, stilistični in slovarski priročnik, do neke mere tudi skladenjski in oblikoslovni. Za naš pravopis je pa značilno, da strokovnjaki jezikoslovni raziskovalci v njem nimajo ravno odločilne besede, ampak v večji meri ostrokovljeni praktiki. Oboji se poleg tega ne prenašajo najbolje, praktiki pa skušajo braniti svoj fevd za vsako ceno. Šele v najnovejšem času imamo pri pravopisu poskus, da bi vendarle prišlo do soglasja med enimi in drugimi, do enotnega nastopanja pred javnostjo, s čimer bi bile zmanjšane možnosti za lov v kalnem, kakor ga ob vsakem pravopisu uspešno uprizarjajo »ljubiteljski« uporabniki jezika in »ljubiteljski« varuhi nikoli dejansko izpričanih in potrjenih jezikovnih vrednot.

Tovrstno stanje, zaznamovano z »bojem« za prestižno vlogo kodifikatorja, je mogoče v informacijski družbi zlahka preseči, saj imamo na voljo objektivizirane

1 V prispevku se z občnoimenskim poimenovanjem slovar slovenskega sodobnega jezika nanašamo na koncept slovarja sodobne slovenščine, podan v Krek et al. (2013b).

jezikovne vire, s katerimi lahko jezikovno podobo uporabniku zelo natančno predstavimo. To pomeni, da se nam ni treba zadovoljiti s tradicionalnimi binarizmi tipa prav – narobe oz. v našem predlogu standardno – nestandardno, temveč lahko jezikovnemu uporabniku poleg tipičnih in pričakovanih informacij o standardnosti posamezne oblike posredujemo tudi informacije o jezikovnem ozadju, drugih relevantnih oblikah ipd., in ravno to je vloga norme v slovarju sodobnega slovenskega jezika – da uporabniku poda čim več informacij o določenem jezikovnem elementu, ne glede na njegov status v trenutno veljavnem jezikovnem predpisu. V predlaganem konceptu so standardizacijske informacije večinoma podane s pomočjo leksikona besednih oblik – Sloleks.[2] Kot izpostavlja Krek (2009: 07), tovrstni leksikoni predstavljajo

> enega od uspelejših srečanj računalniške analize jezika in končnih uporabnikov jezikovnih priročnikov /…/. Predvsem pri slovanskih jezikih z množico pregibnih oblik se uporabniki nemalokrat znajdejo v zadregi glede standardne oblike, te podatke pa potrebujejo tudi sistemi za avtomatsko označevanje in razčlenjevanje besedil. Podobno funkcijo so v predračunalniški dobi opravljali slovarski deli pravopisov in iz obeh so se razvile spletne aplikacije, ki ponujajo tovrstne podatke za končnega uporabnika, temu so navadno dodane tudi druge pravopisne informacije in v nekaterih primerih tudi spletni servisi za pomoč uporabnikom pri jezikovnih vprašanjih.

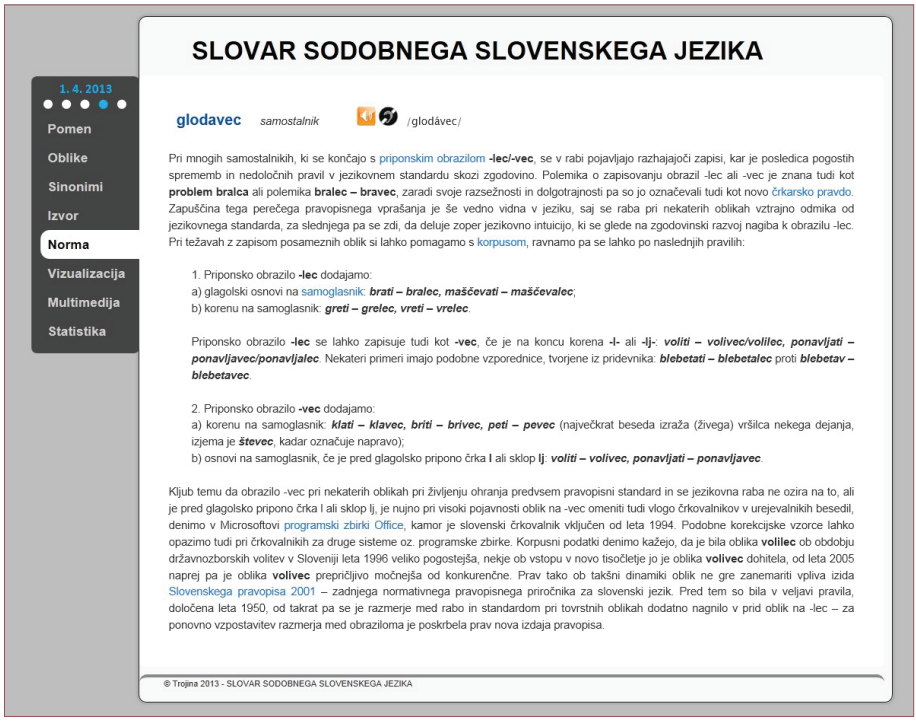


Slika 1: Izpeljane oblike samostalnika *riječ* na Hrvatskem jezičnem portalu.

2 Dostopno na http://www.slovenscina.eu/sloleks (dostop 29. 6. 2015).

To torej pomeni, da lahko v sodobnem, digitalnem in jezikovnotehnološko podprtem slovarskem okolju podatke o standardnosti določene oblike s pomočjo tovrstnih leksikonov zelo hitro pridobimo. Kot primer leksikona za slovanski jezik podajamo informacije o izpeljanih oblikah v okviru geselskega članka Hrvatskega jezičnega portala,[3] ki ga prikazuje Slika 1.

Vprašanje seveda ostaja, kako uporabniku na enem mestu zagotoviti dodatne relevantne informacije o določeni jezikovni zagati. Odgovor na to vprašanje ponuja Slogovni priročnik slovenskega jezika.[4] Pri tem priročniku gre sicer za zunanji vir, podobno kot pri leksikonu Sloleks, vendar ga lahko zaradi združljivosti podatkov zlahka vključimo tudi v druge vire – na ta način nam lahko viri, izdelani v okviru projekta Sporazumevanje v slovenščini, služijo kot skupni ekosistem in zbirališče jezikovnih virov, ki jim je seveda mogoče pridružiti tudi druge vire (gl. Krek et al. 2013a). Predlog za izdelavo Slovarja sodobnega slovenskega jezika tako predvideva tudi neposredno povezavo med posameznimi slovarskimi iztočnicami in vnosi v Slogovnem priročniku, kot prikazuje Slika 2.



Slika 2: Demonstracijski prikaz zavihka Norma v predlaganem Slovarju sodobnega slovenskega jezika (Krek et al. 2013b: 42).

3 Dostopno na http://hjp.novi-liber.hr/index.php?show=kosi_oblici&id=dllgXBY%3D (dostop 22. 7. 2015).
4 Dostopno na http://slogovni.slovenscina.eu (dostop 12. 6. 2015).

Kot lahko vidimo na Sliki 2, norma v predlaganem konceptu predstavlja le del celostne slovarske informacije. Medtem ko lahko uporabnik »gotovost« informacije o standardni varianti (ali standardnih variantah) najde v izhodiščnih zavihkih[5] (pomen in oblika), normativne informacije pojasnjujejo celotno ozadje jezikovne zadrege, in sicer s tem, da mu dajejo na voljo mnoštvo relevantnih jezikovnih informacij in ga privajajo na (zelo realno) dejstvo, da jezik ni monoliten in tog sistem. Seveda tovrstne informacije niso relevantne za vse iztočnice (kjer ne moremo zaznati nobenih zadreg v zvezi z določenim jezikovnim pojavom, nima smisla, da bi poskušali karkoli pojasnjevati) niti za vse uporabnike – če uporabnik nima želje, časa ali volje, da bi se poučil o določeni jezikovni zadregi, bo zavihek in njegovo vsebino pač preskočil in se zadovoljil s tistimi informacijami, ki jih v določeni situaciji potrebuje oz. želi.

Vse to pomeni, da gre pri pripravi normativnih informacij za slovar sodobnega slovenskega jezika v bistvu za pripravo celostnih odgovorov[6] na jezikovne zadrege, zajete v obsežni ontologiji problemskospecifičnih mest v jeziku (gl. Bizjak Končar et al. 2011; H. Dobrovoljc in Krek 2011). Zaradi tega v nadaljevanju nekoliko podrobneje predstavimo portal Slogovni priročnik in njegovo delovanje.

2 ZAKAJ SLOGOVNI PRIROČNIK?

Naslovno vprašanje poglavja je – čeravno je preprosto – v svojem bistvu ambivalentno, in sicer implicira dve različni vprašanji, na kateri odgovarjamo v pričujočem segmentu. Ti vprašanji sta naslednji:

a) Čemu nov priročnik slovenskega jezika?

b) Zakaj ravno poimenovanje slogovni priročnik?

a) Slogovni priročnik slovenskega jezika prinaša nov koncept prosto dostopnega jezikovnega priročnika, ki temelji na korpusnih podatkih in uporabnikom slovenskega jezika omogoča dostop do objektivnih informacij o konkretnih jezikovnih težavah, ne da bi ti za uporabo ali razumevanje potrebovali specialistično jeziko(slo)vno znanje. Portal jim omogoča, da z vnosom neposrednih iskalnih pogojev o določenem problemu pridobijo vse potrebne podatke, da lahko z njimi sami in v relativno kratkem času sprejmejo objektivizirano jezikovno odločitev.

5 V pričujočem prispevku se na posamezne dele predlaganega slovarskega sestavka v digitalnem okolju nanašamo z izrazom zavihek, saj je to pač trenutni kanonski način strukturiranega prikaza informacij v spletnem okolju, obstaja pa seveda velika verjetnost, da se bo to z razvojem spletnih tehnologij korenito spremenilo. Predvsem pričakujemo, da bodo imeli uporabniki spleta v prihodnosti več možnosti, da sami kreirajo obliko in vsebino spletnih strani, kar bo imelo pomembne posledice tudi za spletne slovarje.

6 Poudariti želimo, da gre pri tem res za odgovore, ne rešitve – pri podajanju normativnih informacij namreč nikjer ne želimo implicirati, da je s tovrstnim odgovorom jezikovna zadrega rešena ali odpravljena – je zgolj razložena.

S tem poskušajo tvorci Slogovnega priročnika[7] preseči modaliteto klasičnih priročnikov za slovenski jezik, pri katerih je po navadi treba poznati vsaj nazivno kategorijo za specifično jezikovno zadrego, preden se o njej sploh lahko poučimo. S tovrstnim delovanjem portal v tem oziru deluje ambivalentno, in sicer kot a) pomagalni jezikovni priročnik in b) kot zbirališče jezikovnih virov, kjer so ti tudi eksplicirani ter vizualizirani na različne načine (gl. Krek et al. 2013a), pri tem pa je ena od temeljnih predpostavk avtorjev ta, da

> se je s pojavom spleta, predvsem pa njegove različice 2.0 z množično aktivno participacijo splošne javnosti pri pisanju besedil, ki so poleg tega vsem dostopna takoj (blogi, forumi, družabna omrežja itd.), v temelju spremenila narava in dinamika objavljanja javno dostopnih besedil, kakršna je bila v uporabi večino 20. stoletja. Če je bil predvsem po 2. svetovni vojni vzpostavljen standardni proces objavljanja tiskanih besedil, ki je vključeval avtorja, založbo z urednikom, lektorja in korektorja, se je v zadnjem desetletju, predvsem pa v zadnjih letih z množičnim prehodom najprej v digitalno in potem še v spletno okolje, ta krog v precejšnji meri razdrl (Krek 2012c: 226).

Portal je nastal v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ),[8] v celoti pa se opira na možnosti objektivizirane analize jezika, ki jih prinašajo sodobne jezikovne tehnologije, saj predvsem tuje prakse (prim. npr. Crystal 2006: 257) kažejo, da se v sodobni informacijski družbi ne moremo več zanašati na starodobnike v knjižni obliki, to zavedanje pa je vedno bolj prisotno tudi v domačem prostoru:

> Normiranje jezika je bilo že v obdobjih, ko ta še ni bil povsem ustaljen, težavna naloga, zato je v sodobnem času ob hitro rastočih potrebah družbe, ki zahteva jasen in poljuden, priročen in praktičen jezikovni priročnik, v katerem bo jezikovna informacija predstavljena čim bolj nazorno, postalo očitno, da bo stopnjujoče zahteve uporabnikov jezika mogoče zadovoljiti le, če izkoristimo možnosti in pripomočke, ki jih ta čas ponuja. Jezikovni priročnik sodobnega časa bi zato ob spretni uporabi obstoječih leksikalnih virov in orodij, ki jih omogoča informacijska tehnologija, jezikovnim uporabnikom lahko ponudil odgovore na tista vprašanja, kjer ti dejansko dvomijo in omahujejo (Dobrovoljc in Jakop 2011: 5–6).

Slogovni priročnik deluje predvsem svetovalno in analitično, in sicer z detektiranjem pogostih zadreg v pisnem sporočanju (gl. Bizjak Končar et al. 2011) in ponujanjem celostnih, objektivnih in razumljivih odgovorov nanje, vključno s povezavami na dodatne jezikovne vire. S tem je uresničena ena glavnih teženj snovalcev priročnika – da se uporabnika opremi s prosto dostopnimi, zanesljivimi in predvsem uporabnimi orodji, s katerimi se lahko uspešno udejstvuje v sodobni informacijski družbi, ne da bi se mu bilo treba po pomoč zatekati k lektorjem,

7 Informacije o avtorjih so dostopne na http://www.slovenscina.eu/portali/slogovni-prirocnik (dostop 11. 6. 2015).
8 Rezultati projekta so dostopni na www.slovenscina.eu (dostop 22. 7. 2015).

čeprav je to v slovenski tradiciji povsem običajna praksa (Popič 2014), ki vzpostavlja družbeno neenakost. S tem, ko so govorci prisiljeni delovati v informacijski družbi, ne da bi bili za to primerno (jezikovno) opremljeni, tvegajo, da bodo s tem kršili implicirano normo, kar, kot trdi Škiljan (1999: 185),

> lahko pripelje do družbene diskvalifikacije udeleženca komunikacijskega dejanja, in navsezadnje celo do njegove izključitve iz kanalov javnega komuniciranja. Ta kazen je lahko svoji največkrat implicitni naravi navkljub včasih usodna za posameznika in hkrati kaže na določen tip socialne neenakosti in diferenciacije (na tiste, ki poznajo pravila standardnega jezika/,/ in na tiste, ki jih ne poznajo).

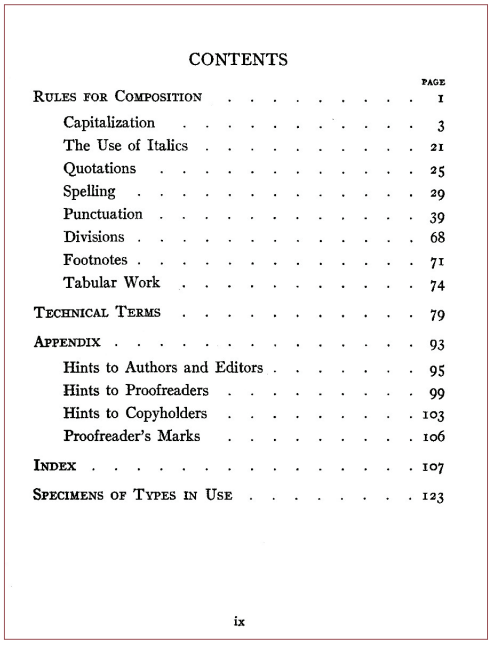
Priročnik torej tako po sestavi kot po modaliteti podajanja informacij predstavlja nov pristop v slovenskem prostoru, njegova prihodnja vloga v procesu kodifikacije slovenskega jezika pa je odvisna predvsem od (jezikovno)političnih odločitev, to pa izhaja predvsem iz nejasnega in v preteklosti pogosto tudi arbitrarnega procesa kodifikacije slovenščine (Krek 2012č; Krek 2014).

Ob opredelitvi vloge priročnika v slovenskem kodifikacijskem okolju se poraja vprašanje, zakaj se priročnik – tehnološki podstati navkljub – opira in nanaša na druge kodifikacijske priročnike. Korpusni in drugi podatki, pridobljeni pri ostalih aktivnostih projekta SSJ, so bili namreč dovolj celoviti, da bi v marsikaterem segmentu Slogovni priročnik že lahko utemeljeno presegel zastarel pravopisni predpis ali slovnično pravilo. Razlog za to je ta, da so se tvorci priročnika odločili za model, ki se ves čas nanaša na jezikovni standard, kakršen pač v določenem trenutku je. Tako priročnik aktualnemu predpisu venomer nastavlja rabno zrcalo in kaže variantnost, kot jo kažejo podatki iz sodobne javne pisne produkcije. Uporabnik priročnika lahko tako hitro razbere, kakšna je kodifikacija v razmerju do rabe, obenem pa spozna tudi, ali je kodifikacija v določenem primeru smiselno zastavljena ali pa jo je glede na rabo bolj smiselno kršiti in doseči želeni komunikacijski učinek.

Lastnost, ki Slogovni priročnik nemara najbolj ločuje od tradicionalnih pravopisnih priročnikov v slovenskem prostoru, je ta, da portal nima nobene ambicije, da bi jezik spreminjal. V tem pogledu je zgolj registrator stanja v jeziku, največ, kar si v tej vlogi dovoli, je interpretacija rabe, kodifikacije in odnosa med njima, vse ostale informacije predstavljajo preverljivi podatki. Pomembna prednost Slogovnega priročnika je tudi ta, da uporabnika poučuje o jezikovnih zakonitostih in ga na ta način osvobodi jezikovnih avtoritet in mu daje suverenost pri jezikovnih odločitvah, hkrati pa ga opremi tudi z dovolj podatki, da lahko to odločitev tudi sprejme.

b) Samo poimenovanje slogovni priročnik se razlikuje od tradicionalnih poimenovanj za jezikovne priročnike v slovenskem prostoru, npr. brus, antibarbarus, pravopis, pravipis, rešeto ipd. Razlog za to je predvsem ta, da ni namenjen kakršni koli

obliki purifikacije slovenskega jezika, temveč obveščanju jezikovnega uporabnika in seznanjanju le-tega z objektiviziranimi podatki o splošni rabi. Izraz izhaja iz založništva, naslanja pa se na standardizacijo pisnega izražanja znotraj neke organizacije ali širše sfere,[9] vendarle pa nosi (družbeni) ugled tudi zunaj teh okvirov – slogovni priročnik neke organizacije lahko namreč služi kot priročnik za pisanje v številnih drugih organizacijah, ki se v večji ali manjši meri ukvarjajo z založniško ipd. dejavnostjo. Nekateri tovrstni priročniki lahko tako postanejo standard za specifičen del besedilotvorja, npr. za citiranje, oblikovanje prispevkov ipd. Za ponazoritev razlik in podobnosti med slovensko tradicijo pravopisja in angloameriško tradicijo, ki je nosilka tradicije slogovnih priročnikov, v nadaljevanju podajamo popis vsebine prvega slovenskega pravopisa (Levec 1899) in pa prve izdaje slogovnega priročnika Chicago Manual of Style (1906),[10] ki sta izšla v istem časovnem obdobju. V slogovnem priročniku za angleščino tako najdemo naslednje (vsebinske) kategorije:



CONTENTS

ix

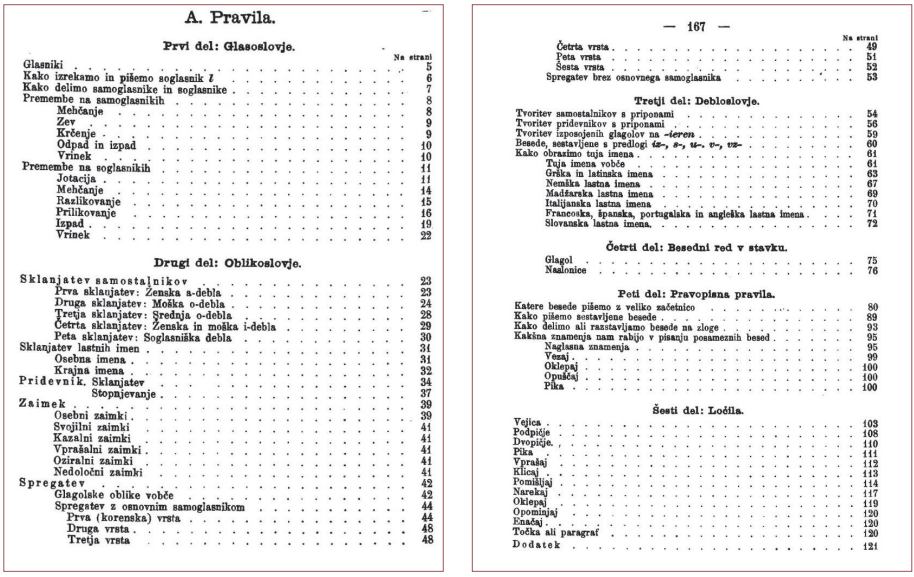
Slika 3: Vsebina prve izdaje priročnika Chicago Manual of Style.

Prva izdaja priročnika Chicago Manual of Style je torej zajemala naslednje vsebinske sklope: velika in mala začetnica, raba ležečega tiska, citiranje, črkovanje, ločila, deljenje besed, sprotne opombe, tabele, terminologija, nasveti za avtorje in urednike, nasveti za korektorje, nasveti za založnike, korekturna znamenja.

9 Prim. npr. Medinstitucionalni slogovni priročnik za inštitucije Evropske unije, ki »/v/sebuje enotna slogovna pravila in konvencije, ki jih morajo upoštevati vse institucije, organi in agencije Evropske unije«, tudi za slovenski jezik. Dostopno na http://publications.europa.eu/code/sl/sl-000100.htm (dostop 20. 7. 2015).

10 Priročnik Chicago Manual of Style je obenem tudi primer priročnika s širšim krogom veljave in vpliva.

Vsebina v prvem slovenskem pravopisnem priročniku je precej drugačna, kot lahko vidimo na Sliki 4 v nadaljevanju:



A. Pravila.

Prvi del: Glasoslovje.

Drugi del: Oblikoslovje.

Tretji del: Besloslovje.

Četrti del: Besedni red v stavku.

Peti del: Pravopisna pravila.

Šesti del: Ločila.

Slika 4: Kazalo vsebine SP 1899.

Kot razkrije že bežen pogled na kazali vsebine, je slovenski priročnik ne le podrobneje segmentiran, temveč zajema bistveno več kot le standardizacijo pisne norme, saj vsebuje tudi glasoslovne, oblikoslovne in druge podatke; vsebuje torej slovnico kot tudi slovar.[11] To smer so ubrali vsi nadaljnji slovenski pravopisi, tako da je, kot pravi Dobrovoljc (2004: 108–109),

> slovenski pravopis v tem trenutku veliko več, kot pove njegovo ime, ki je v tem smislu tudi nekoliko zavajajoče: je kompleksni normativni priročnik, ki podaja sistemski pregled izrazne podobe jezika (tudi na osnovi pomena) in rešitve pogostih jezikovnih težav. Pri nas namreč nimamo tako imenovanega specializiranega pravopisnega priročnika, kot jih imajo nekateri drugi jeziki /.../ ali stroke.

V tem je torej bistvena razlika med slogovnim priročnikom in pravopisom ob aplikaciji na slovensko okolje: medtem ko poskuša biti pravopis priročnik tipa »vse v enem«, obenem pa temelji na predpisovalnem načelu usmerjanja jezikovne rabe, je temeljno vodilo slogovnega priročnika to, da uporabnika seznani z naborom pisnih možnosti, ki jih ima glede na podatke iz splošne rabe v določenem

11 Po drugi strani pa seveda drži, da tradicionalne slovnice v angloameriški tradiciji zajemajo tudi pravopisna pravila. Zanimivo je predvsem to, da se te v samem predpisu zaradi različnih jezikoslovnih teorij, ki jim avtorji sledijo, lahko tudi razlikujejo (Quirk et al. 1985: 13).

trenutku na voljo in jih opremi z interpretacijo glede na veljavni standard – vse dodatne informacije o določenem jezikovnem problemu pa seveda služijo kot razlaga določene jezikovne zadrege.

Druga pomembna razlika je ažuriranost jezikovnega priročnika, ki pa ne izhaja (nujno) iz njegove tehnološke ali netehnološke zasnove. Kot prikaz drugačnega razumevanje jezikovnih sprememb bomo uporabili ista priročnika, le da bomo tokrat uporabili sezname sprememb od predzadnje do zadnje izdaje, in sicer bomo tako na kratko ponazorili spremembe med 5. (1997) in 6. izdajo Slovenskega pravopisa (2001) ter spremembe, ki so bile sprejete med 15. (2003) in 16. izdajo priročnika Chicago Manual of Style (2010). V nadaljevanju podajamo popis sprememb v slogovnem priročniku za angleščino:[12]

- obširne posodobitve v skladu s sodobno rabo, tehnologijo in poklicno prakso;
- razširjen pregled nad elektronskim založništvom, vključno s postopki za korigiranje spletnih in drugih elektronskih dokumentov;
- razširjeno poglavje o nepristranskem izražanju (politična korektnost);
- razširjeno poglavje o pravični rabi in avtorskih pravicah elektronskih besedil;
- razširjeno poglavje o paralelnih strukturah v povedi (medstavčno ujemanje v skladenjskih kategorijah);
- nov in izboljšan vodnik za rabo vezaja v preprostem, tabelaričnem prikazu;
- uvod v sistem unikod, mednarodni računalniški standard za zapisovanje črk in simbolov za pisanje v različnih jezikih; vključena je tudi tabela s šiframi sistema unikod;
- dodatni sklici na slogovne priročnike in citatne standarde drugih organizacij;
- posodobljeni napotki za uporabo naslovov DOI v razmerju do naslovov URL, z več primeri;
- več nasvetov, kako citirati bloge, poddaje[13] in druge elektronske vire;
- poenostavljeni nasveti, kako citirati pravne in upravne dokumente;
- temeljito revidirani postopki izdelave dokumentov, vključno s pregledom elektronskega označevanja in jezika XML;
- posodobljen glosar z obširnejšo terminologijo, povezano z elektronskim založništvom;
- razširjena poglavja za številčenje odstavkov in več navzkrižnih referenc za lažje brskanje po priročniku, zlasti na spletu;
- jasnejši nasveti in pravila.

12 Povzeto po spletni strani http://www.chicagomanualofstyle.org/about16.html (dostop 20. 7. 2015).
13 Ang. *podcast*.

Kot lahko vidimo, se večina sprememb nanaša na (elektronsko) založništvo, citatne standarde, predvsem v elektronskem (založniškem) okolju, ter delo z digitalnimi tehnologijami. Za primerjavo v nadaljevanju podajamo še povzetek sprememb v Slovenskem pravopisu med 5. in 6., zadnjo izdajo (povzeto po Majcenovič 2001):

- ustreznejša formulacija nekaterih pravil (npr. Z veliko je priporočljivo pisati svojilne pridevnike iz lastnih imen tudi tedaj, kadar zaznamujejo duhovno last … → Z veliko se svojilni pridevniki iz lastnih imen lahko pišejo tudi tedaj, kadar zaznamujejo duhovno last …);
- dopolnitev nekaterih pravil (npr. V teoloških in bogoslužnih besedilih je dovoljeno pisati tudi S*veti Duh*.);
- spremembe zgledov (spremembe tehničnega tipa, nadomestila z zgledi, ki so za ponazoritev pravila boljši, dodajanje zgledov, črtanje zgledov, preverjanje nejezikovne ustreznosti);
- spremembe v razdelkih Glasoslovje in Pisave za posamezne jezike;
- dodana Slovaropisna pravila;
- spremembe v slovarju.

Čeprav je med izidoma primerjanih priročnikov skorajda desetletje razlike, kar je v informacijski dobi razmeroma dolgo obdobje,[14] lahko ob primerjavi sprememb brez težav vidimo konceptualne razlike v dojemanju življenjskih okolij jezika in z njimi povezanih jezikovnih dejstev v obeh okoljih oz. pri obeh načinih jezikovne standardizacije – ameriški priročnik je v sedmih letih od izida prejšnje različice pri snovanju novih pravil zaznal obilico novosti v (elektronskem) besedilotvornem postopku, medtem ko je slovensko pravopisje ostalo zvesto tradicionalističnemu konceptu posodabljanja pravil, ne glede na to, da so se ravno v tem obdobju odvijale spremembe, ki so bile za proces standardizacije in vlogo jezikovnih priročnikov temeljne in nespregledljive.

Če torej poskušamo odgovoriti na vprašanje, zakaj Slogovni priročnik – ne gre za priročnik, ki bi se ukvarjal s stilistiko, temveč gre za celostni jezikovni priročnik, ki odgovarja na akutna mesta v slovenskem jeziku na vseh jezikovnih ravneh, ob tem pa uporabnikom ponuja takojšnjo jezikovno pomoč in jim omogoča uspešno udejstvovanje v sodobni informacijski družbi. S tem se v določeni meri navezuje na (angloameriško) tradicijo slogovnih priročnikov, poimenovanje pa prekinja navezavo na to, kar se v slovenskem okolju tradicionalno dojema kot pravopis. Preden pa portal Slogovni priročnik predstavimo nekoliko podrobneje, v sledečem poglavju predstavimo nekaj tujih zgledov jezikovnih priročnikov v digitalnem okolju.

14 Ni pa odveč še enkrat poudariti, da so spremembe, uvedene v 6. izdaji pravopisnih pravil, tudi *zadnje*, gre torej za najbolj ažurirano različico slovenske kodifikacije.

3 TUJI ZGLEDI

V pričujočem poglavju predstavimo nekaj tujih zgledov iz evropskih okolij, v katerih so se jezikovni priročniki preselili v digitalno okolje (informacije o ureditvah v nemškem, danskem, nizozemskem in češkem okolju so na voljo tudi v Krek (2012č) in Bizjak Končar et al. 2011). Digitalnost je sicer splošna smernica, ki velja tudi za okolja s podobno normativno ureditvijo kot pri nas (Bizjak Končar et al. 2011: 17). V osnovi lahko koncepte, ki temeljijo na sodobnih tehnologijah, ločujemo glede na temeljno usmeritev, in sicer lahko ti a) jemljejo sodobne tehnologije zgolj kot medij za diseminacijo informacij o jezikovnem standardu ali pa b) privzamejo sodobne jezikovne tehnologije kot osnovo procesa standardizacije, s katerimi ne le diseminiramo jezikovni standard, temveč ga tudi osnujemo, analiziramo in verificiramo.[15]

Tipičen predstavnik standardizacijskega priročnika, ki izrablja spletne tehnologije kot dodatni medij, je denimo hrvaški portal Jezični savjeti,[16] ki ga prikazuje Slika 5.

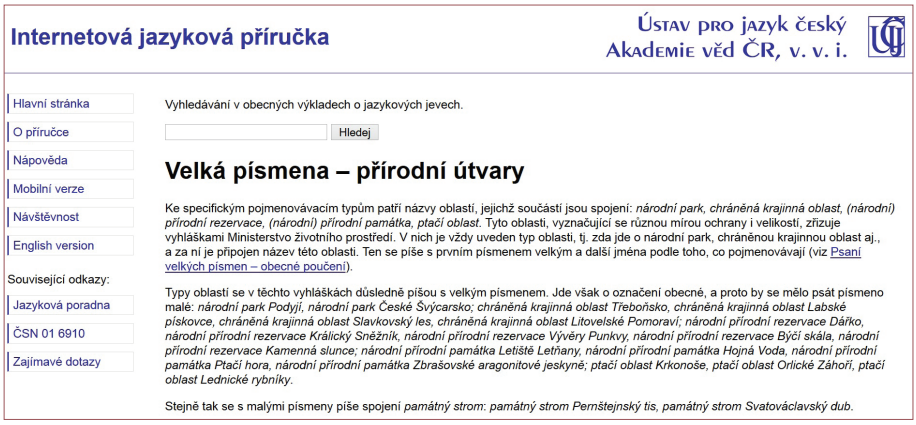
Slika 5: Razlaga pravopisnega pravila na spletnem portalu Jezični savjeti.

Pri tem portalu gre za jezikovno svetovalnico, podobno svetovalnici Inštituta za slovenski jezik Frana Ramovša, ki se nanaša predvsem na eksplicitacijo ali razlago

15 Trenutnega slovenskega standardizacijskega priročnika ne moremo zlahka uvrstiti v nobeno od skupin, saj gre pri pravilskem delu zgolj za digitalno reprodukcijo knjižne izdaje; je pa seveda veliko bližje prvi skupini. V tem pogledu je slovenska standardizacija podobna slovaški (prim. Pravidlá slovenského pravopisu – dostopno na http://www.juls.savba.sk/ediela/psp2000/psp.pdf; dostop 22. 7. 2015).

16 Dostopno na http://savjetnik.ihjj.hr/ (dostop 22. 7. 2015).

pravil hrvaškega pravopisa. Podobno deluje tudi spletišče Internetová jazyková příručka za češki jezik,[17] ki deluje kot spletni pravopis – tukaj torej ne gre zgolj za eksplicitacijo ali razlago pravil, temveč so pravila podana v obliki, prilagojeni spletnemu okolju, kot prikazuje Slika 6:



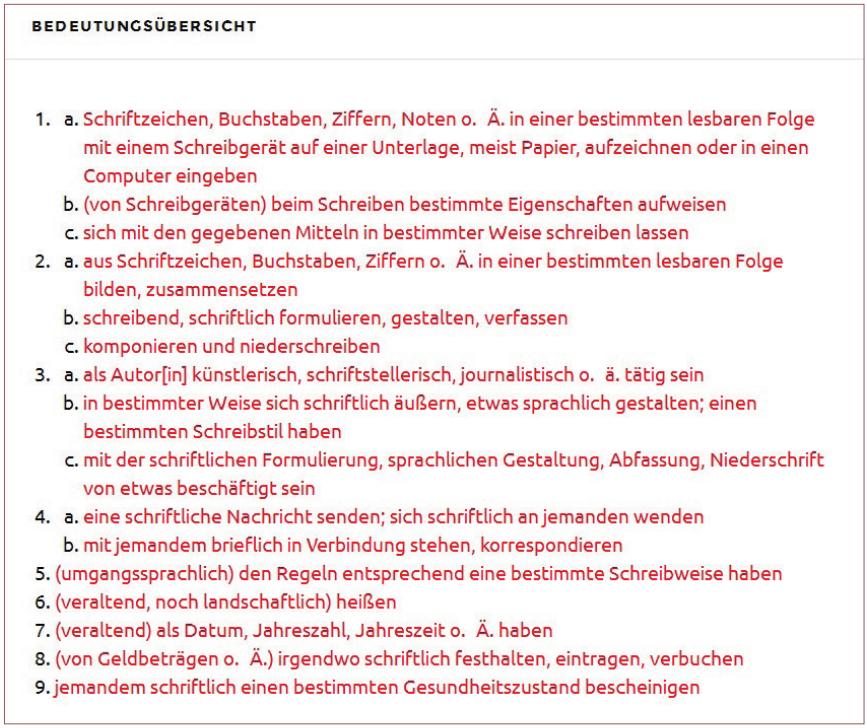
Slika 6: Internetová jazyková příručka.

V nadaljevanju nekoliko podrobneje predstavimo enega od nemških jezikovnih priročnikov, in sicer Dudnov spletni portal,[18] predvsem zaradi tega, ker lahko med slovensko in nemško jezikovno ureditvijo najdemo številne vzporednice, poleg tega pa je imelo nemško kulturno okolje, s tem pa tudi jezikovno, na slovensko velik vpliv.[19] Kljub temu da gre pri nemški v osnovi za direktivno ureditev, saj znotraj nemško govorečega okolja deluje pravopisni svet,[20] je ta direktivnost teht(a)na – gl. npr. priporočila iz decembra 2010, ki temeljijo na korpusnojezikoslovnih raziskavah sodelavcev založb Duden in Wahrig.[21] V teh priporočilih tako pravopisni svet predlaga, da se določene variante črtajo, nekatere pa sprejmejo, gre pa za primere, »pri katerih je v pisni rabi prišlo do premika preferenc«. Dudnov spletni priročnik izpostavljamo tudi zato, ker gre za primer (učinkovitega) zbiranja jezikovnih virov in ponazarjanja raznorodnih jezikovnih podatkov iz obsežnega Dudnovega ekosistema (podobno kot je tudi ena od ambicij Slogovnega priročnika, da učinkovito združi informacije iz različnih segmentov ekosistema Sporazumevanje v slovenščini in drugih virov).[22] Slika 7 prikazuje začetek slovarskega gesla za nemški glagol *schreiben* ('pisati') na Dudnovem portalu:

17 Dostopno na http://prirucka.ujc.cas.cz/ (dostop 22. 7. 2015)
18 Dostopno na http://www.duden.de/ (dostop 22. 9. 2014).
19 Za nemško govorno okolje obstaja tudi spletni portal Grammis, gre pa za skladenjski informacijski sistem Inštituta za nemški jezik. Portal je dostopen na http://hypermedia.ids-mannheim.de/ (dostop 22. 7. 2015), podrobneje pa je predstavljen v Bizjak Končar et al. (2011: 22–25).
20 Rat für deutsche Rechtschreibung (dostopno na http://www.rechtschreibrat.com/; dostop 30. 7. 2015).
21 Dostopno na http://rechtschreibrat.ids-mannheim.de/download/empfehlungen2010.pdf (dostop 30. 7. 2015).
22 O Slogovnem priročniku kot zbirališču in torišču jezikovnih virov pišejo Krek et al. (2013b).

Slika 7: Osnovni pravopisni podatki o iztočnici *schreiben*.

Iztočnično geslo je na Dudnovem portalu predstavljeno večnivojsko, in kot lahko vidimo na primeru na Sliki 7, osnovni (pravopisni) segment vsebuje podatke o besedni vrsti, konjugaciji, pogostosti pojavljanja, deljenju besede in primere kolokacij. Slika 8 prikazuje pomenski del iztočničniškega članka:

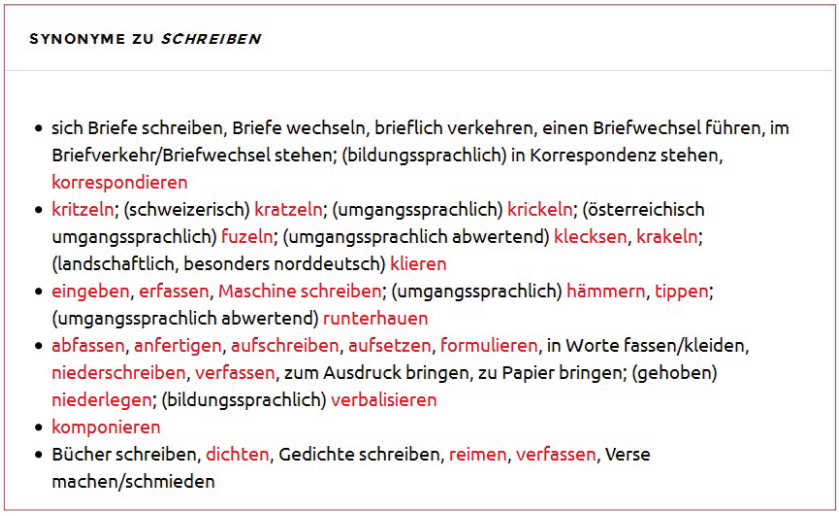


Slika 8: Pomenski del gesla.

Takoj za pomenskim delom gesla, ki je razvejan in zaradi možnosti, ki jih ponuja digitalni medij, tudi razširjen in pregleden, sledi sekcija »Ste vedeli?«, ki zajema različna opozorila, napotke ipd., s katero želijo redaktorji uporabnika opozoriti na specifike, vezane na določen jezikovni element. Slika 9 tako prikazuje opozorilo, da gre pri glagolu *schreiben* za eno od »besed«, ki jih je treba poznati za pridobitev certifikata o znanju nemščine:[23]

Slika 9: Opozorilo, da je beseda del osnovnega besedišča za pridobitev certifikata.

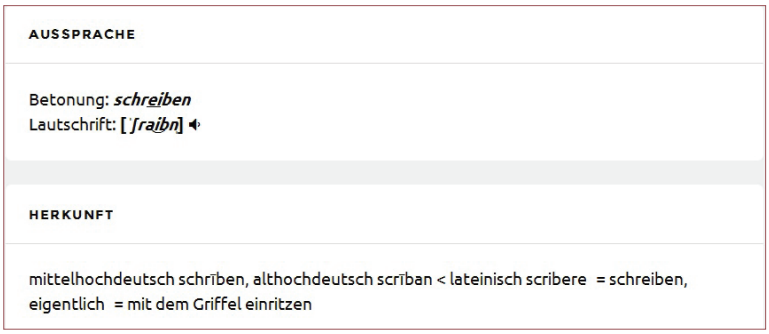
Sklopu z opozorili sledijo sopomenke, ki so na ta način integrirane v del glavnega gesla. Kot so pokazale študije specializiranih uporabnikov (gl. Čibej et al. 2015), so informacije o sopomenkah eden najbolj zaželenih in iskanih segmentov, ki pa slovenski leksikografiji še manjka. Slika 10 prikazuje obravnavo sopomenk glagola *schreiben* v Dudnovem priročniku:



Slika 10: Sopomenke za glagol *schreiben*.

Za sopomenkami sledi sklop, ki na preprost način podaja izgovarjavo iztočnice, in sicer je naglašeni zlog preprosto podčrtan, temu delu pa sta dodana fonetični zapis in pa izgovarjava. Kot prikazuje Slika 11, informacijam o izgovarjavi sledijo informacije o izvoru iztočnice:

23 Dostopno na http://www.goethe.de/lrn/prj/pba/bes/gzb/deindex.htm?wt_sc=b1 (dostop 5. 9. 2015).



AUSSPRACHE

Betonung: ***schreiben***
Lautschrift: [***ʃraibn***]

HERKUNFT

mittelhochdeutsch schrīben, althochdeutsch scrīban < lateinisch scribere = schreiben, eigentlich = mit dem Griffel einritzen

Slika 11: Izgovorjava glagola *schreiben* in etimološke informacije.

Izgovarjavi in etimološkim informacijam sledi pregibalni vzorec, v primeru na Sliki 12 z dopolnilom o tem, ali gre za pravilni ali nepravilni glagol ter napotilom, s katerim pomožnim glagolom se glagol *schreiben* veže v perfektu.

GRAMMATIK

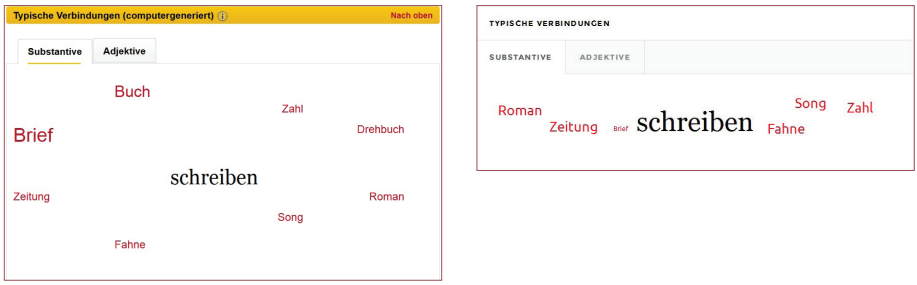
starkes Verb; Perfektbildung mit »hat«

PRÄSENS	INDIKATIV	KONJUNKTIV I	IMPERATIV
SINGULAR	ich schreibe	ich schreibe	
	du schreibst	du schreibest	schreib, schreibe!
	er/sie/es schreibt	er/sie/es schreibe	
PLURAL	wir schreiben	wir schreiben	

Slika 12: Standardni pregibni vzorec za glagol *schreiben*.

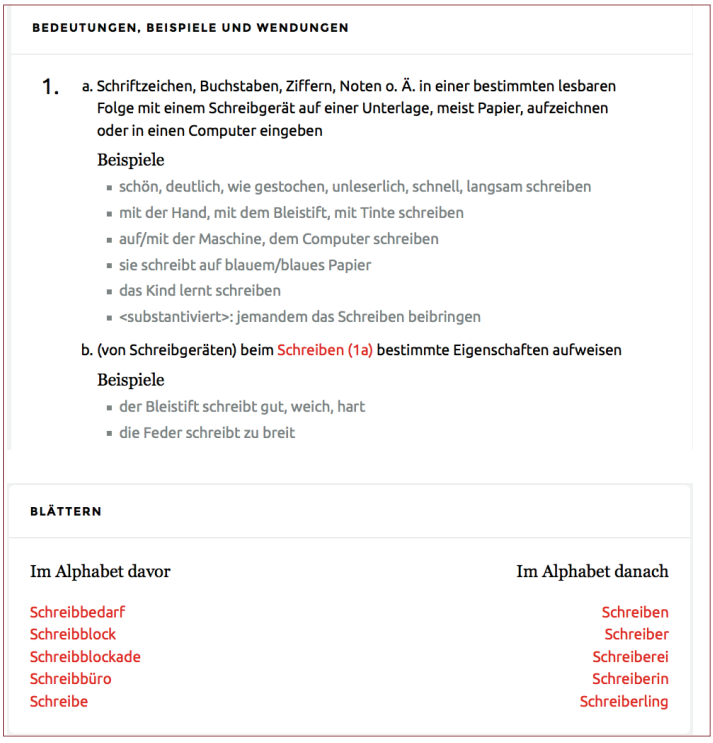
Izjemno zanimiv segment predstavlja odsek Značilne zveze, ki zajema samodejno generirano predstavitev najpogostejših kolokatorjev za iztočnico, kot prikazuje Slika 13:[24]

24 Ob koncu julija 2015 je bil Dudnov portal prenovljen. Med prenovo je poleg grafičnih in strukturnih sprememb umanjkala tudi navedba pri pogostih zvezah, da gre za računalniško pridobljene oz. generirane podatke, kar je dober pokazatelj spremenjene optike v leksikografiji – avtomatizirano pridobivanje jezikovnih podatkov ni več razumljeno kot posebnost, temveč kot integralen del leksikografske dejavnosti.



Slika 13: Samodejno generirani kolokatorji za glagol *schreiben* pred prenovo portala (levo) in po njej (desno).

Kot lahko vidimo, so kolokati predstavljeni nazorno in razumljivo, prav tako pa so logično dimenzionirani – pogostejši kolokatorji so zapisani v večji pisavi kot manj pogosti. Temu segmentu sledi še sklepni del s pomeni in primeri ter sekcija »za radovedne«, kjer lahko uporabniki izberejo katero od sosednjih gesel po abecedi, kot prikazuje Slika 14:



Slika 14: Pomeni s primeri in sklepni del »za radovedne« – najbližja gesla po abecedi.

Dudnov spletni portal nam lahko služi kot zgled sodobnega jezikovnega priročnika predvsem zaradi naslednjih vidikov:

- gre za obsežen in celovit prikaz različnih in raznovrstnih informacij iz obsežnega Dudnovega ekosistema;
- uporabnik lahko na eni strani izve veliko o »življenjskih vzorcih« iskane besede, in sicer na različnih jezikovnih ravninah;
- informacije so podane na razumljiv in pregleden način, uporabniku pa ni treba poznati jezikovnosistemskih tipologij, da bi lahko informacije razumel;
- priročnik poskuša animirati in pritegniti uporabnika, da bi ta iz lastnega zanimanja nadaljeval z brskanjem po slovarju.

Naštete poudarke oz. primere dobre prakse bi veljalo vključevati tudi v prihodnje leksikografske dejavnosti v slovenskem kulturnem okolju, v veliki meri pa so bila ravno ta načela ključna za izdelavo Slogovnega priročnika slovenskega jezika, ki ga predstavljamo v sledečem poglavju.

4 SLOGOVNI PRIROČNIK SLOVENSKEGA JEZIKA

V pričujočem poglavju predstavimo vsebinski in tehnični del Slogovnega priročnika slovenskega jezika, obenem pa se osredotočimo tudi na način implementacije informacij iz priročnika v slovarju sodobnega slovenskega jezika.

4.1 Nastanek

Slogovni priročnik slovenskega jezika je nastal v okviru projekta Sporazumevanje v slovenščini, aktivnosti pri izdelavi pa so potekale med januarjem 2010 in decembrom 2013.[25] Namen aktivnosti je bil

> detektirati težavne točke pri pisanju besedil in analizirati dosedanje reakcije tvorcev besedil na jezikovno rabo in normativne priročnike za slovenščino. Namen je analiza možnosti prestavljanja težišča jezikovne kompetence pri sestavljanju besedil s specializiranih strokovnjakov za jezik na jezikovnotehnološko sodobno opremljenega tvorca besedil (Bizjak Končar et al. 2011: 4).

To pomeni, da je bil cilj aktivnosti predvsem opolnomočiti jezikovnega uporabnika, da bi ta lahko sam brez vsakršne pomoči sprejemal jezikovne odločitve in v končni fazi – uspešno tvoril besedila v slovenskem jeziku. Seveda tovrstnega

25 Informacije o nastanku in procesu izdelave Slogovnega priročnika povzete po Bizjak Končar et al. (2011).

premika ne bi bilo mogoče doseči z drastično drugačnim procesom izdelave priročnika, ki je moral v celoti temeljiti na objektiviziranih jezikovnih podatkih, obenem pa bi moral biti v končni različici prosto dostopen na spletu. Na to so opozorili že zgledi iz tujih okolij, tudi takšnih s podobnimi normativnimi ureditvami (gl. Bizjak Končar et al. 2011: 17–25).

Izdelava Slogovnega priročnika je potekala v več etapah, in sicer je bilo najprej treba ugotoviti, kje imajo uporabniki najpogosteje težave pri pisanju:

> Pri detektiranju uporabnikovih težav in potreb smo združili oba postopka, ki so ju do sedaj uporabljali avtorji t. i. izproblemskih jezikovnih priročnikov, in sicer analizo jezikovne produkcije in zbiranje vprašanj. Vendar pa se je način analize in zbiranja podatkov močno spremenil, saj nam razvoj jezikovnih tehnologij omogoča vsaj deloma avtomatizirano analizo bistveno obsežnejšega gradiva, posplošitve pa so zato mnogo zanesljivejše (Bizjak Končar et al. 2011: 9).

Vzporedno z analizo gradiva je bila vzpostavljena obsežna tipologija oz. ontologija jezikovnih zadreg (več informacij o oblikovanju ontologije v H. Dobrovoljc in Krek 2011). Ta tipologija je seveda nekončna množica, nanaša pa se na posamezne jezikovne ravnine in jezikovne segmente, tako da jo je mogoče dopolnjevati glede na potrebe in sočasen razvoj jezika. Ko je bila ontologija glede na trenutne potrebe jezikovnih uporabnikov dopolnjena, je bilo mogoče začeti z vsebinskim popolnjevanjem pripravljenega okvira. V trenutni, testni fazi Slogovni priročnik zajema 15 celostnih odgovorov na jezikovne zadrege.[26]

4.2 Vsebina

Portal Slogovnega priročnika sestavlja več med seboj povezanih delov. Vsebina sloni na seznamu 700 jezikovnih zadreg. Te jezikovne zadrege so[27]

> bile detektirane pri analizi treh različnih tipov primarnih virov: tradicionalnih pravopisnih priročnikov, besedilnih korpusov in spletnih forumov z jezikovnimi vprašanji. Spisek zadreg je sestavljen kot ontološko drevo, na vrhu katerega je osem osnovnih kategorij: PRAVOPIS (A), PRAVOREČJE (B), OBLIKOSLOVJE (C), BESEDOTVORJE (D), BESEDIŠČE (E), SKLADNJA (F), BESEDILO (G), RAZNO (H). Te vrhnje kategorije se cepijo na podkategorije, trenutno najdlje do šestega nivoja, za označevanje kategorij pa je bil izbran sistem kombinacij črk in številk, ki omogoča poljubno zrnavost pri različnih kategorijah – vrhnje kategorije se na podkategorije namreč cepijo neenakomerno (Krek 2012č: 227).

26 Demonstracijski zgledi so na voljo na spletni strani http://slogovni.slovenscina.eu/Kazalo/Kazalo (dostop 17. 7. 2015).
27 Gl. tudi Dobrovoljc in Krek (2011).

Posamezne odgovore sestavljajo kratki odgovor, dolgi odgovor ter dodatna rubrika za navdušence s povezavami na druge digitalne in digitalizirane vire.

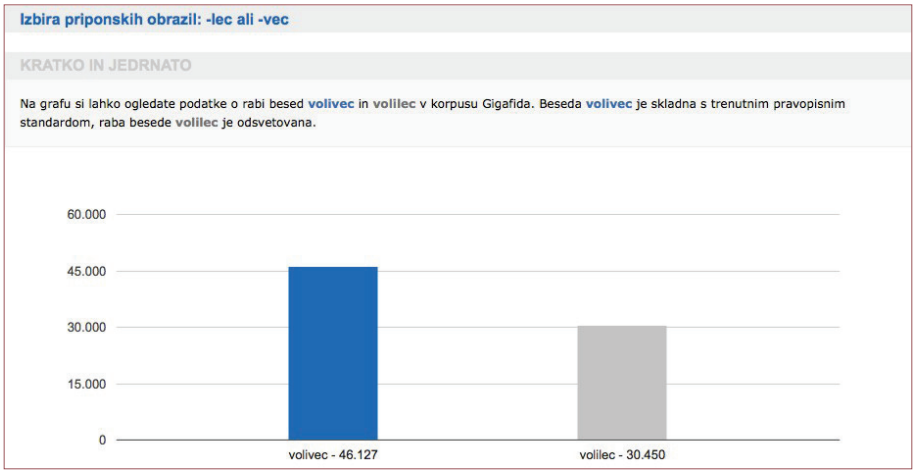
4.2.1 Kratki odgovori

Kratki odgovor

> sestavlja besedilo v formatu XML, ki v stavčni obliki zagotavlja ustrezen izpis statističnih podatkov iz besedilnega korpusa Gigafida in iz leksikona besednih oblik. Zasnovan je kot univerzalni mehanizem, s katerim je mogoče opisati statistično stanje pri vseh možnih kombinacijah standardnih ali nestandardnih oblik pri določeni kategoriji (Krek 2012č: 227).

Eden ključnih elementov sistema je ta, da je kratki odgovor, ki se odraža v grafu (tam, kjer je odgovor pač možno podati z grafom), dinamičen in lahko vedno odseva realno stanje v jeziku. Ob vsakršni posodobitvi besedilnega korpusa, ki služi kot podstat slogovnega priročnika, se morebitne spremembe v jeziku, ki jih korpus registrira, takoj prikažejo na grafu. V določenih primerih, pri katerih je koristno, da si uporabnik ogleda celotno paradigmo pregibnih oblik, je v kratkem odgovoru tudi povezava na leksikon Sloleks.

Slika 15 prikazuje graf kratkega odgovora za problem izbire priponskega obrazila -lec ali -vec. Kot lahko vidimo, se razmerje trenutno nekoliko nagiba proti obliki volivec. Kratki odgovor prinaša tudi kratko informacijo o tem, kaj o dani problematiki pravi trenutna kodifikacija slovenskega jezika (standardne variante se na grafu tudi vidno ločijo od ostalih).[28]



Slika 15: Kratki odgovor v Slogovnem priročniku.

28 Podatki o reprezentaciji standardnega v pričujočem poglavju so deloma prirejeni po Arhar Holdt et al. (2013).

Kratki odgovor za isto jezikovno zadrego je v formatu XML zapisan takole:

```
<odgovor_Kratko id=«D2c1«>
<tabela>
<!-- lema: »belivec«, »volivec«, »brivec«, »hvastavec« ... -->
<beseda katera=«1« tip=«lema« povezava=«gigafida«>
     <p_oblika att=«SPSP« val=«D2c1«/>
     <pogoj att=«norma« val=«/|nestandardno|nejasno«/>
</beseda>
<!-- lema: »belilec«, »volilec«, »brilec«, »hvastalec« ... -->
<beseda katera=«2« tip=«lema« povezava=«gigafida«>
     <p_oblika att=«SPSP« val=«D2c1«/>
     <pogoj att=«norma« val=«/|nestandardno|nejasno«/>
</beseda>
</tabela>

<!-- varianta DVA, standardno2 /belilec/belivec/-->
<tekst var=«S00.S00« graf=«12«>Na grafu si lahko ogledate podatke
o rabi besed <beseda katera=«1«/> in <beseda katera=«2«/> v korpusu Gigafida. Obliki sta variantni, raba obeh je skladna s trenutnim
pravopisnim standardom.
</tekst>

<!-- varianta DVA, standardno1, nestandardno2, >50% /volilec/
volivec/-->
<tekst var=«S02.N01« graf=«12«>Na grafu si lahko ogledate podatke o rabi besed <beseda katera=«1«/> in <beseda katera=«2«/> v
korpusu Gigafida. Beseda <beseda katera=«1«/> je skladna s trenutnim pravopisnim standardom. Raba besede <beseda katera=«2«/>
je odsvetovana, vendar tega navodila večina piscev ne upošteva.
</tekst>

<!-- varianta DVA, standardno1, nestandardno2, <50% /volilec/
volivec/-->
<tekst var=«S01.N02« graf=«12«>Na grafu si lahko ogledate podatke o rabi besed <beseda katera=«1«/> in <beseda katera=«2«/> v
korpusu Gigafida. Beseda <beseda katera=«1«/> je skladna s trenutnim pravopisnim standardom, raba besede <beseda katera=«2«/>
je odsvetovana.
</tekst>

<!-- varianta DVA, nejasno1, nejasno2 /hvastalec/hvastavec/ -->
<tekst var=«J00.J00« graf=«12«>Na grafu si lahko ogledate podatke o rabi besed <beseda katera=«1«/> in <beseda katera=«2«/> v
korpusu Gigafida. Skladnost obeh variant s pravopisnim standardom
je odvisna od interpretacije. Za podrobnejše pojasnilo si oglejte
odgovor »na dolgo in široko«.
</tekst>

<!-- varianta EN, standardno1 /brivec/ -->
<tekst var=«S00« graf=«1«>Na grafu si lahko ogledate podatke o
rabi besede <beseda katera=«1«/> v korpusu Gigafida. Beseda je
skladna s trenutnim pravopisnim standardom.
</tekst>
```

```
<!-- varianta EN, standardno2 /brivec/ -->
<tekst var=«S00« graf=«2«>Na grafu si lahko ogledate podatke o
rabi besede <beseda katera=«2«/> v korpusu Gigafida. Beseda je
skladna s trenutnim pravopisnim standardom.
</tekst>

<!-- varianta EN, standardno4 /brilec/ -->
<tekst var=«s00.N00« graf=«12«>
Na grafu si lahko ogledate podatke o rabi besede <beseda kate-
ra=«2«/> v korpusu Gigafida. Raba besede je odsvetovana, vendar
tega navodila pisci ne upoštevajo. Oblika, ki je po pravopisnem
standardu priporočena, je <beseda katera=«1«/>, vendar se v kor-
pusu Gigafida ne pojavlja.
</tekst>

</odgovor_Kratko>
```

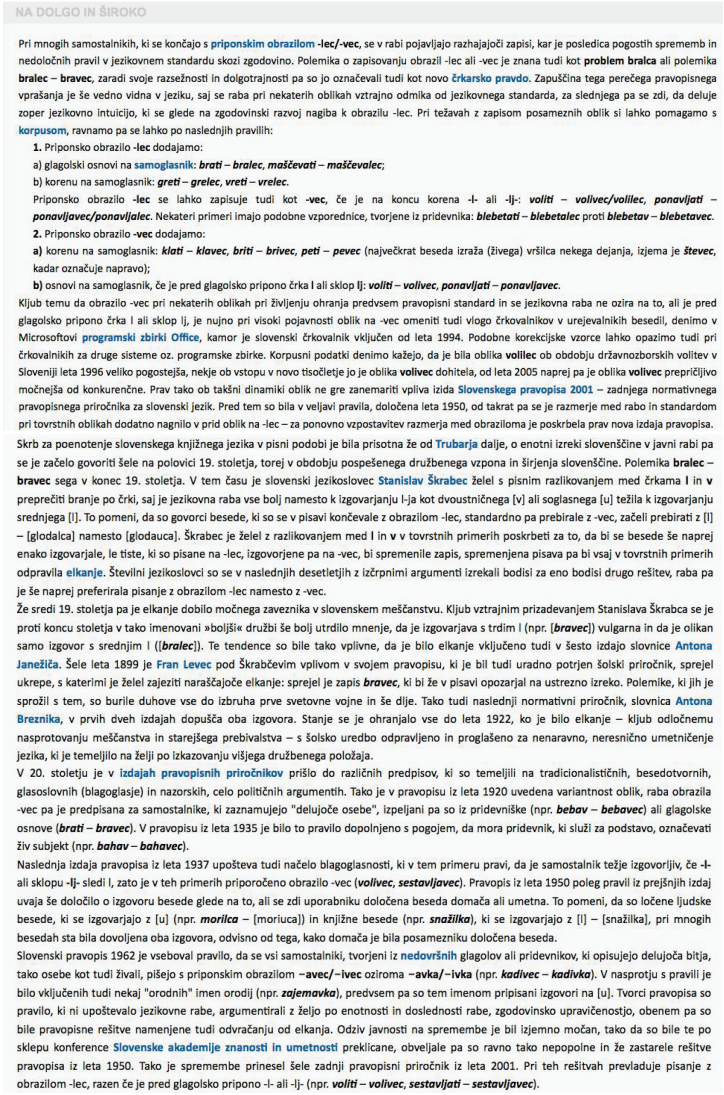
Kot lahko vidimo v zapisu XML, je za vsako normativno zadrego predviden nabor relevantnih zapisovalnih možnosti oz. možnosti, ki se pojavljajo v rabi. Te možnosti so ovrednotene na podlagi trenutno veljavnega pravopisnega standarda, od tega pa je odvisno, kako so predstavljene na grafu v okviru kratkega odgovora.

4.2.2 Dolgi odgovor

Za razliko od kratkega je dolgi odgovor, ki je zapisan v standardnem formatu HTML, statičen, kar pomeni, da ga lahko osvežujemo le z ročnim spreminjanjem HTML-kode. Dolgi odgovor v osnovi podaja naslednje informacije oz. podatke:

- podatke o veljavnem jezikovnem standardu;
- podatke o rabi;
- sistemske opredelitve razmerja standard – raba;
- povezave;
  - povezave na zunanje vire;
  - poljudne razlage strokovnih terminov, rabljenih v dolgem odgovoru;
  - povezave do spiskov besed, ki spadajo v isto problemsko kategorijo (in so v leksikonu besednih oblik označene z isto oznako);
- jezikovnosistemske podatke, ki pojasnjujejo ozadje jezikovne zadrege (kjer je to mogoče);
- etimološke in zgodovinskoslovnične osvetlitve;
- morebitna opozorila;
- problemskospecifične informacije.

Namen dolgega odgovora je torej ta, da poda podrobnejše, strokovne in argumentirane informacije o danem jezikovnem problemu, istočasno pa uporabnika opremi še z informacijami o drugih dostopnih virih, ki so mu lahko kakor koli v pomoč. Slika 7 kaže dolgi odgovor na zadrego s pisanjem obrazila *-vec*/*-lec* – kot lahko vidimo pri odgovoru, ta odgovarja na celotno paradigmo oz. na posamezno kategorijo v ontologiji.[29]



NA DOLGO IN ŠIROKO

Pri mnogih samostalnikih, ki se končajo s **priponskim obrazilom -lec/-vec**, se v rabi pojavljajo razhajajoči zapisi, kar je posledica pogostih sprememb in nedoločnih pravil v jezikovnem standardu skozi zgodovino. Polemika o zapisovanju obrazil -lec ali -vec je znana tudi kot **problem bralca** ali polemika **bralec – bravec**, zaradi svoje razsežnosti in dolgotrajnosti pa so jo označevali tudi kot novo **črkarsko pravdo**. Zapuščina tega perečega pravopisnega vprašanja je še vedno vidna v jeziku, saj se raba pri nekaterih oblikah vztrajno odmika od jezikovnega standarda, za slednjega pa se zdi, da deluje zoper jezikovno intuicijo, ki se glede na zgodovinski razvoj nagiba k obrazilu -lec. Pri težavah z zapisom posameznih oblik si lahko pomagamo s **korpusom**, ravnamo pa se lahko po naslednjih pravilih:

1. Priponsko obrazilo **-lec** dodajamo:

a) glagolski osnovi na **samoglasnik**: ***brati – bralec, maščevati – maščevalec***;

b) korenu na samoglasnik: ***greti – grelec, vreti – vrelec***.

Priponsko obrazilo **-lec** se lahko zapisuje tudi kot **-vec**, če je na koncu korena **-l-** ali **-lj-**: ***voliti – volivec/volilec, ponavljati – ponavljavec/ponavljalec***. Nekateri primeri imajo podobne vzporednice, tvorjene iz pridevnika: ***blebetati – blebetalec*** proti ***blebetav – blebetavec***.

2. Priponsko obrazilo **-vec** dodajamo:

a) korenu na samoglasnik: ***klati – klavec, briti – brivec, peti – pevec*** (največkrat beseda izraža (živega) vršilca nekega dejanja, izjema je ***števec***, kadar označuje napravo);

b) osnovi na samoglasnik, če je pred glagolsko pripono črka **l** ali sklop **lj**: ***voliti – volivec, ponavljati – ponavljavec***.

Kljub temu da obrazilo -vec pri nekaterih oblikah pri življenju ohranja predvsem pravopisni standard in se jezikovna raba ne ozira na to, ali je pred glagolsko pripono črka l ali sklop lj, je nujno pri visoki pojavnosti oblik na -vec omeniti tudi vlogo črkovalnikov v urejevalnikih besedil, denimo v Microsoftovi **programski zbirki Office**, kamor je slovenski črkovalnik vključen od leta 1994. Podobne korekcijske vzorce lahko opazimo tudi pri črkovalnikih za druge sisteme oz. programske zbirke. Korpusni podatki denimo kažejo, da je bila oblika **volilec** ob obdobju državnozborskih volitev v Sloveniji leta 1996 veliko pogostejša, nekje ob vstopu v novo tisočletje jo je oblika **volivec** dohitela, od leta 2005 naprej pa je oblika **volivec** prepričljivo močnejša od konkurenčne. Prav tako ob takšni dinamiki oblik ne gre zanemariti vpliva izida **Slovenskega pravopisa 2001** – zadnjega normativnega pravopisnega priročnika za slovenski jezik. Pred tem so bila v veljavi pravila, določena leta 1950, od takrat pa se je razmerje med rabo in standardom pri tovrstnih oblikah dodatno nagnilo v prid oblik na -lec – za ponovno vzpostavitev razmerja med obrazilома je poskrbela prav nova izdaja pravopisa.

Skrb za poenotenje slovenskega knjižnega jezika v pisni podobi je bila prisotna že od **Trubarja** dalje, o enotni izreki slovenščine v javni rabi pa se je začelo govoriti šele na polovici 19. stoletja, torej v obdobju pospešenega družbenega vzpona in širjenja slovenščine. Polemika **bralec – bravec** sega v konec 19. stoletja. V tem času je slovenski jezikoslovec **Stanislav Škrabec** želel s pisnim razlikovanjem med črkama **l** in **v** preprečiti branje po črki, saj je jezikovna raba vse bolj namesto k izgovarjanju l-ja kot dvoustničnega [v] ali soglasnega [u] težila k izgovarjanju srednjega [l]. To pomeni, da so govorci besede, ki so se v pisavi končevale z obrazilom -lec, standardno pa prebirale z -vec, začeli prebirati z [l] – [glodalca] namesto [glodauca]. Škrabec je želel z razlikovanjem med **l** in **v** v tovrstnih primerih poskrbeti za to, da bi se besede še naprej enako izgovarjale, le tiste, ki so pisane na -lec, izgovorjene pa na -vec, bi spremenile zapis, spremenjena pisava pa bi vsaj v tovrstnih primerih odpravila **elkanje**. Številni jezikoslovci so se v naslednjih desetletjih z izčrpnimi argumenti izrekali bodisi za eno bodisi drugo rešitev, raba pa je še naprej preferirala pisanje z obrazilom -lec namesto z -vec.

Že sredi 19. stoletja pa je elkanje dobilo močnega zaveznika v slovenskem meščanstvu. Kljub vztrajnim prizadevanjem Stanislava Škrabca se je proti koncu stoletja v tako imenovani »boljši« družbi še bolj utrdilo mnenje, da je izgovarjava s trdim l (npr. [***bravec***]) vulgarna in da je olikan samo izgovor s srednjim l ([***bralec***]). Te tendence so bile tako vplivne, da je bilo elkanje vključeno tudi v šesto izdajo slovnice **Antona Janežiča**. Šele leta 1899 je **Fran Levec** pod Škrabčevim vplivom v svojem pravopisu, ki je bil tudi uradno potrjen šolski priročnik, sprejel ukrepe, s katerimi je želel zajeziti naraščajoče elkanje: sprejel je zapis ***bravec***, ki bi bi že v pisavi opozarjal na ustrezno izreko. Polemike, ki jih je sprožil s tem, so burile duhove vse do izbruha prve svetovne vojne in še dlje. Tako tudi naslednji normativni priročnik, slovnica **Antona Breznika,** v prvih dveh izdajah dopušča oba izgovora. Stanje se je ohranjalo vse do leta 1922, ko je bilo elkanje – kljub odločnemu nasprotovanju meščanstva in starejšega prebivalstva – s šolsko uredbo odpravljeno in proglašeno za nenaravno, neresnično umetničenje jezika, ki je temeljilo na želji po izkazovanju višjega družbenega položaja.

V 20. stoletju je v **izdajah pravopisnih priročnikov** prišlo do različnih predpisov, ki so temeljili na tradicionalističnih, besedotvornih, glasoslovnih (blagoglasje) in nazorskih, celo političnih argumentih. Tako je v pravopisu iz leta 1920 uvedena variantnost oblik, raba obrazila -vec pa je predpisana za samostalnike, ki zaznamujejo "delujoče osebe", izpeljani pa so iz pridevniške (npr. ***bebav – bebavec***) ali glagolske osnove (***brati – bravec***). V pravopisu iz leta 1935 je bilo to pravilo dopolnjeno s pogojem, da mora pridevnik, ki služi za podstavo, označevati živ subjekt (npr. ***bahav – bahavec***).

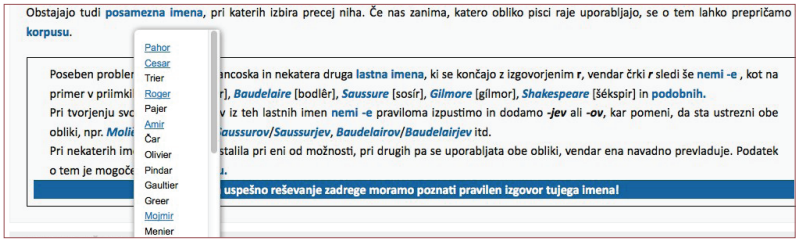
Naslednja izdaja pravopisa iz leta 1937 upošteva tudi načelo blagoglasnosti, ki v tem primeru pravi, da je samostalnik težje izgovorljiv, če **-l-** ali sklopu **-lj-** sledi l, zato je v teh primerih priporočeno obrazilo -vec (***volivec, sestavljavec***). Pravopis iz leta 1950 poleg pravil iz prejšnjih izdaj uvaja še določilo o izgovoru besede glede na to, ali se zdi uporabniku določena beseda domača ali umetna. To pomeni, da so ločene ljudske besede, ki se izgovarjajo z [u] (npr. ***morilca*** – [moriuca]) in knjižne besede (npr. ***snažilka***), ki se izgovarjajo z [l] – [snažilka], pri mnogih besedah sta bila dovoljena oba izgovora, odvisno od tega, kako domača je bila posamezniku določena beseda.

Slovenski pravopis 1962 je vseboval pravilo, da se vsi samostalniki, tvorjeni iz **nedovršnih** glagolov ali pridevnikov, ki opisujejo delujoča bitja, tako osebe kot tudi živali, pišejo s priponskim obrazilom **–avec/–ivec** oziroma **–avka/–ivka** (npr. ***kadivec – kadivka***). V nasprotju s pravili je bilo vključenih tudi nekaj "orodnih" imen orodij (npr. ***zajemavka***), predvsem pa so tem imenom pripisani izgovori na [u]. Tvorci pravopisa so pravilo, ki ni upoštevalo jezikovne rabe, argumentirali z željo po enotnosti in doslednosti rabe, zgodovinsko upravičenostjo, obenem pa so bile pravopisne rešitve namenjene tudi odvračanju od elkanja. Odziv javnosti na spremembe je bil izjemno močan, tako da so bile te po sklepu konference **Slovenske akademije znanosti in umetnosti** preklicane, obveljale pa so ravno tako nepopolne in že zastarele rešitve pravopisa iz leta 1950. Tako je spremembe prinesel šele zadnji pravopisni priročnik iz leta 2001. Pri teh rešitvah prevladuje pisanje z obrazilom -lec, razen če je pred glagolsko pripono -l- ali -lj- (npr. ***voliti – volivec, sestavljati – sestavljavec***).

Slika 16: Dolgi odgovor v Slogovnem priročniku.

29 Odgovor je neposredno dostopen prek povezave http://slogovni.slovenscina.eu/Search/SearchStringDolgoInSiroko?q=D2c1 (dostop 22. 7. 2015).

Del dolgega odgovora so tudi paradigmatski spiski – pri teh gre za nabor primerov, ki spadajo v isto kategorijo kot konkretna iskalna poizvedba.[30] Na ta način lahko uporabniku na enem mestu podamo celostno informacijo o določeni paradigmi, in če je določen jezikovni element povezan z leksikonom Sloleks, lahko uporabnik na ta primer preprosto klikne in brskalnik ga samodejno preusmeri na leksikon, kjer se lahko pouči o pregibalnih vzorcih za iskano poizvedbo.



Slika 17: Paradigmatski seznam imen na govorjeni *r*.

Tovrstne spiske lahko izdelamo s samodejnimi postopki, vendar pa jih je treba pred končnim pregledom še validirati, kar lahko napravimo z množičenjem.[31] Postopek množičenja je za izdelavo paradigmatskih spiskov v Slogovnem priročniku izjemno priročen, saj lahko na ta način hitreje obdelamo zajetne količine podatkov in izločimo šum. Seveda to ne pomeni, da s tem množičarjem prepuščamo zapletene in/ali interpretativne odločitve, temveč jim zastavljamo zelo jasna in konkretna, enodimenzionalna vprašanja o posameznih jezikovnih elementih, kot prikazuje Slika 18:

Slika 18: Navodilo za množičenje pri izbiri priponskega obrazila *-lec* ali *-vec* za Slogovni priročnik.

30 Podrobneje o luščenju podatkov za Slogovni priročnik pišeta K. Dobrovoljc in Krek (2013).
31 Ang. crowdsourcing. Postopek, vezan na tovrstno leksikografsko dejavnost, je obširno predstavljen v Fišer et al. (2015).

V Tabeli 1 si lahko ogledamo vmesne rezultate glasovanja o konkretni jezikovni zadregi (izbiri priponskega obrazila *-lec* ali *-vec*):

Tabela 1: Rezultati množičenja pri odločanju o veljavnosti obrazila *-vec/-lec*.

Koren	Beseda	Con-firmed	Next	Rejec-ted	Skupna vsota	Konč-na
SP_D2c1_1	volivec - volilec	22		2	24	DA
SP_D2c1_10	glodavec - glodalec	19		2	21	DA
SP_D2c1_100	plenivec - plenilec	3		1	4	?
SP_D2c1_101	sejavec - sejalec	2		2	4	?
SP_D2c1_102	bravec - bralec	3		2	5	?
SP_D2c1_103	ubijavec - ubijalec	3		2	5	?
SP_D2c1_104	dajavec - dajalec	4		1	5	DA
SP_D2c1_105	stavec - stalec	1	1	3	5	?
SP_D2c1_106	vrevec - vrelec	2	1	2	5	?
SP_D2c1_107	izdajavec - izdajalec	3		1	4	?
SP_D2c1_108	iglavec - iglalec	1		3	4	?
SP_D2c1_109	plazivec - plazilec	4		1	5	DA
SP_D2c1_11	delivec - delilec	24		10	34	DA
SP_D2c1_110	kmetovavec - kmetovalec	1	1	2	4	?
SP_D2c1_111	skakavec - skakalec	4		1	5	DA
SP_D2c1_112	zakonodajavec - zakonodajalec	5			5	DA
SP_D2c1_113	kozovec - kozolec	4	1		5	DA
SP_D2c1_114	igravec - igralec	4	1		5	DA
SP_D2c1_115	morivec - morilec	2		3	5	?

Šele z uporabo obsežnih zbirk (izluščenih) podatkov, ki so po potrebi očiščeni še v postopku množičenja, lahko uporabniku zagotovimo objektivne podatke o normativnem ozadju določenega jezikovnega elementa. S tem lahko podamo zelo natančno sliko o tem, kakšna je trenutna jezikovna raba – tudi pri mestih, ki so z vidika trenutnega jezikovnega standarda povsem dvojnična. Slika 19 kot primer celostnih podatkov prikazuje eno od tovrstnih dvojničnih mest v slovenskem jeziku – izbira variantne podaljšave pri angleških in francoskih imenih na govorjeni *r*.

koren	Frekvence v korpusu Gigafida koren+rejev [rejev, rejevega, rejevemu …] npr. **Shakespearejev**	koren+reov [reov, reovega, reovemu …] npr. **Shakespeareov**	koren+rjev [rjev, rjevega, rjevemu …] npr. **Shakespearjev**
Shakespea	27	32	699
Gilmo	69	7	28
Pha	5	5	
Ha	30	10	1
Car	18		
Tou	15		1
Tor	8		
Bar	11	2	1
Guer	6		
Kompa	6		
tola	5		
Nu	529		
g	1		
Do	17		
D	404	4	
Kada	10		30
Commodo	6		59
Rove	10		357
Da	91		6
Magui	3	5	
Netwa		4	
Lefebv		4	
Moo	3	38	

Slika 19: Variantnost podaljšave imen z *-rev/rjev/-rov/-reov*.

Kot lahko vidimo na Sliki 19, variantnost nikakor ni inherentna celotni paradigmi, temveč variira med posameznimi primeri in posameznimi variantami. Vse to so pri odločanju o izbiri določene variante pomembni podatki, ki jih lahko s pomočjo Slogovnega priročnika uporabniku tudi podamo na preprost in pregleden način.

4.2.3 Rubrika »Za navdušence«

Zadnji del odgovora kaže na zunanje povezave, predvsem na referenčne vire, kjer lahko uporabnik dobi natančnejše podatke ne le o trenutnem predpisu, temveč tudi o zgodovini kodifikacije dane jezikovne zadrege. Uporabnik je pri tem usmerjen na spletno različico Slovenskega pravopisa 2001, digitalizirane pravopise 1899–1962 ter starejše slovnične priročnike. Povezave se bodo dopolnjevale, če oziroma ko bodo na voljo novi priročniki ali pa bodo obstoječi priročniki prešli v javno domeno, obenem pa bodo dodani tudi drugi viri – znanstveni članki, jezikovni kotički, sporočila Jezikovnega razsodišča ipd., tako da s tem odgovor v okviru priročnika postane celostno zbirališče informacij o določeni jezikovni zadregi. Slika 20 v nadaljevanju prikazuje rubriko Za navdušence za izbrano jezikovno zadrego:

Slika 20: Rubrika Za navdušence v Slogovnem priročniku.

4.3 Normativne informacije v Slogovnem priročniku

Slogovni priročnik podaja normativne informacije na dva načina: prvi je posreden, in sicer z navajanjem jezikovnega standarda znotraj dolgega odgovora, drugi pa je imanentna povezava med priročnikom in normativnimi podatki v leksikonu besednih oblik Sloleks.[32] Format Lexical Markup Framework (LMF), ki je bil uporabljen pri sestavljanju leksikona, namreč omogoča, da je vsaki leksikonski enoti oz. obliki mogoče pripisati poljubno število dodatnih informacij, npr. izgovor, normativni status itd., kot prikazuje Slika 21 (prevzeto iz Krek et al. 2013b: 390; relevantni odseki kode so obarvani):

```
<LexicalEntry id="LE_S_Matija" xmlns:d="urn:LEKSIKON_SSJ">
     <feat att="besedna_vrsta" val="samostalnik" />
     <feat att="vrsta" val="lastno_ime" />
     <feat att="spol" val="moški" />
     <feat att="SPSP" val="C1a2a" />
     <Lemma>
          feat att="zapis_oblike" val="Matija" />
     </Lemma>
<…>
     <WordForm>
     <feat att="število" val="ednina" />
     <feat att="sklon" val="rodilnik" />
          <FormRepresentation>
               <feat att="zapis_oblike" val="Matija" />
               <feat att="msd" val="Slmer" />
               <feat att="SPSP" val="C1a2a" />
               <feat att="norma" val="variantno" />
               <feat att="tip" val="C1a2a_s_1" />
               <feat att="pogostnost" val="858" />
          </FormRepresentation>
          <FormRepresentation>
               <feat att="zapis_oblike" val="Matije" />
               <feat att="msd" val="Slmer" />
               <feat att="SPSP" val="C1a2a" />
               <feat att="norma" val="variantno" />
               <feat att="tip" val="C1a2a_s_2" />
               <feat att="pogostnost" val="4018" />
          </FormRepresentation>
     </WordForm>
<…>
</LexicalEntry>
```

Slika 21: Normativne leksikonske informacije v formatu LMF.

32 Informacije o reprezentaciji standardnega in nestandardnega v virih SSJ so na voljo v Arhar Holdt et al. (2013).

Posamezna enota v leksikonu besednih oblik postane del portala Slogovni priročnik šele takrat, ko ji je pripisana določena kategorija iz nabora jezikovnih zadreg, kar pomeni, da je za portal enota brez pripisa kategorije nevidna oz. nerelevantna. Normativne kategorije so zaradi občutljivosti normativnih podatkov in jezikoslovčeve presoje pripisane ročno, analiza korpusnih podatkov, ki vodi do odločitev o pripisu kategorije, pa je v veliki meri avtomatizirana.

5 SKLEP

Slovenski jezik je bil v zadnjem času postavljen pred digitalne izzive, ki jim slovenistična stroka ni znala (pogosto tudi želela) slediti, to pa se kaže tudi v pomanjkanju sodobnih jezikovnih orodij oziroma posluha zanje – obstoječi pravopisni priročniki preprosto ne detektirajo problemov, ki jih imajo govorci/pišoči v spremenjenem digitalnem okolju tvorjenja in sprejemanja besedil. Hitri tempo življenja in pospešena digitalizacija v globalnem merilu širita prepad tako med jeziki, ki so na sodobne izzive pripravljeni, in jeziki, ki se s temi izzivi ne znajo ali (v strahu pred globalizacijskim razvrednotenjem) ne želijo spoprijeti, obenem pa se širi tudi prepad med digitalnimi domačini in tistimi, ki se z novimi tehnologijami ne znajdejo najbolje – kar hitro vodi v t. i. informacijsko revščino, tj. stanje neenakopravnosti na podlagi slabšega dostopa do informacij.

Slogovni priročnik slovenskega jezika je namenjen vsem uporabnikom slovenskega jezika, ki se znajdejo pred jezikovno zadrego in potrebujejo takojšen odgovor; rečemo pa lahko, da je priročnik po svoji zasnovi še posebno prilagojen mlajšim generacijam in sodobni besedilni produkciji. Dobljene informacije so objektivne, povedne in se ne zanašajo na intuicijo, za vse, ki jih obravnavana problematika bolj zanima, pa so podane tudi podrobnejše informacije kot tudi zunanji viri. Namen je torej opolnomočiti jezikovnega uporabnika, da bo lahko sam uspešno (so)deloval v sodobni informacijski družbi, obenem pa je cilj tudi, da se razvrednoti prestižni položaj slovenskega kodifikatorja – in to izključno z dostopnimi, odprtimi in transparentnimi podatki, na katerih Slogovni priročnik tudi temelji in so lahko (ali pa bi celo morali biti) sestavni del slovarja sodobnega slovenskega jezika.

III
Slovarski uporabniki

Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi koraki

Špela Arhar Holdt

Abstract
This paper provides a reflection on the state and role of dictionary user research in Europe, with particular focus on the situation in Slovenia. Over fifty years of research have been conducted abroad into dictionary use and user needs, expectations and abilities. While lexicographers in Slovenia often discuss and refer to (potential) dictionary users and their needs, these discussions have not yet been supported by empirical data. Definitions of the various user types, i.e. target, general, common, average and demanding, remain generalised, lacking substance and internal structure. As such, they are prone to reflecting the ideas, opinions and observations of those participating in the discussion. In this paper, the importance of user research in Slovenian lexicography is emphasised. A diagram is presented of dictionary use and different user groups, which may serve not only as a basis for further discussion on dictionary users but also as a starting point for systematic research into their needs. In the second part of the paper, the possibilities for including dictionary users in the dictionary creation process are presented, and a number of user-oriented tasks are suggested for inclusion in the plans for a new monolingual dictionary of Slovenian.

Keywords: Slovenian lexicography, monolingual dictionary, user research, user groups, user participation

Ključne besede: slovenska leksikografija, enojezični slovarji, uporabniške raziskave, uporabniške skupine, sodelovanje uporabnikov

1 UVOD

Raziskave slovarske rabe in slovarskih uporabnikov imajo v evropskem prostoru dolgo tradicijo, čeprav so se v primerjavi z drugimi vrstami slovarskih raziskav pojavile relativno pozno. Razvoj področja uporabniških raziskav lahko spremljamo od prvih pobud v šestdesetih letih prejšnjega stoletja (npr. v Barnhart 1962; Householder 1967), prek osemdesetih, ki jih imamo lahko za obdobje utemeljevanja (npr. Tomaszczyk 1979; Hartmann 1987; Wiegand 1987), devetdesetih, ki prinesejo razcvet na področju slovaropisja za potrebe učenja angleščine kot drugega/tujega jezika (npr. Atkins 1998; Nesi 2000; Tono 2001), vse do preloma v novo tisočletje, ki z vstopom slovarja v digitalni svet ponudi nove metodološke možnosti (kar opisujejo npr. De Schryver 2003; Tarp 2009; Müller Spitzer 2014).

Tekom desetletij so bili za raziskovanje slovarske rabe in slovarskih uporabnikov preizkušeni različni metodološki postopki, od anketiranja in intervjujev, raziskovanja uporabe slovarjev (npr. s pomočjo uporabniških popisov slovarske rabe ali z metodo uporabniškega »glasnega razmišljanja«, sledenja pogledu (angl. *eye tracking*), analiz dnevniških datotek (angl. *log file analysis*) ali s pomočjo poizvedb o uporabnikih v samem slovarskem vmesniku), do vodenih eksperimentov oz. testov in uporabniških evalvacij posameznih jezikovnih priročnikov.[1] Naštete metode uporabniških raziskav so precej različne, vendar je vsem skupno raziskovalno izhodišče: raziskave slovarske rabe in slovarskih uporabnikov so nujna podpora obstoječim slovaropisnim projektom in neizbežen predpogoj za razvoj slovaropisja in njegovih izdelkov. Slovarji so lahko različnih vrst, različnega poslanstva, z različnim ciljnim naslovnikom, vendar vse združuje dejstvo, da so namenjeni (tudi in predvsem) uporabi. S tega stališča se zdi povsem samoumevno, da se slovaropisje uporabnikom raziskovalno posveča in ciljno, kontinuirano spremlja njihove potrebe, želje in zmožnosti. In res se področje uporabniških raziskav ireverzibilno postavlja ob bok drugim vsebinam, ki so slovaropisce zanimala skozi zgodovino. Nekatere zgodnejše uporabniške raziskave, predvsem ankete, so bile v zadnjem času sicer kritično ovrednotene (Bogaards 2003; Tarp 2009), pa tudi pri nekaterih sodobnih raziskavah je mogoče identificirati določene pomanjkljivosti (Bergenholtz 2011), vendar je kvaliteta in kvantiteta raziskav v porastu.[2] V zadnjih letih je v literaturi denimo opaziti napredne analize dnevniških datotek (npr. Bergenholtz in Johnsen 2005, De Schryver et al. 2006, Verlinde in Binon 2010, Wolfer et al. 2012, Trap Jensen et al. 2014) in rezultate obsežnejših anketiranj slovarskih uporabnikov (npr. Lorentzen in Theilgaard 2012, Müller-Spitzer 2014). Na drugi strani je opaziti razvoj funkcijske teorije slovaropisja (angl. *function theory of lexicography*)

1 S. Tarp našteva: anketiranje, intervjuvanje, opazovanje rabe, izvedbo zapisov o rabi slovarja (ang. *dictionary protocols*), eksperimentiranje, testiranje in analizo dnevnikov iskanj (Tarp 2009: 283).

2 Trenutno morda najbolj pereči problem je tematska razdrobljenost področja, na kar opozarja denimo (Müller-Spitzer 2014: 19).

(Fuertes-Olivera in Tarp 2014), ki slovarskega uporabnika razume kot smisel in posledično nepogrešljivega partnerja v razvoju slovaropisja:

> Če jih razumemo kot pripomoček, so slovaropisni izdelki po temeljnih značilnostih enaki kateremu koli drugemu človeškemu orodju, zasnovani so namreč – oz. bi morali biti – za zadovoljevanje določene vrste človeških potreb. Slovaropisni izdelki vedno predstavljajo odnos med vsaj dvema osebama, uporabnikom in slovaropiscem; kot taki so slovarski pripomočki v bistvu družbena stvaritev in slovaropisje je družbena veda (Fuertes-Olivera in Tarp 2014: 45, prevod Š. A. H.).

Čeprav se zdi skoraj nemogoče, da bi raziskovalni vrvež hitro razvijajočega se področja uporabniških raziskav v slovaropisnih krogih lahko ostal neopažen, za slovenski prostor trenutno velja prav slednje. Posamezne razprave sicer pozivajo k empiričnemu pristopu k slovarskim uporabnikom (npr. Stabej 2009; Logar 2009), vendar v stroki še vedno vlada diskurz, v katerem se o (potencialnih) slovarskih uporabnikih sicer razmišlja in nanje sklicuje, vendar diskusija še ni podprta z (domačimi ali tujimi) objektivnimi podatki iz uporabniške prakse.[3] Pri tem se uporabljajo koncepti, kot so *splošni*, *običajni, povprečni, zahtevni* uporabnik, ki ostajajo pomensko nedefinirani in notranje nestrukturirani, s tem pa izpostavljeni nevarnosti, da postanejo element za preslikavo idej, mnenj in opažanj razpravljavca samega. Stanje je torej še vedno primerljivo s problemom, ki ga je pred tridesetimi leti identificiral G. Hatherall: če predpostavljamo, da se pripravljavci slovarja trudijo izdelati priročnik, ki bo za uporabnike kolikor je mogoče uporaben, se je treba vprašati tudi, kako pripravljavci slovarja sploh *vedo,* kaj uporabnik zares potrebuje. Avtor odgovarja, da pripravljavci slovarja teh podatkov nimajo, ampak se pri delu opirajo na predvidevanja, odločitve pa sprejemajo na podlagi lastne uporabniške izkušnje (Hatherall 1984: 183). Zato želimo v tej monografiji kot alternativo obstoječemu stanju: opredeliti načelno držo do slovarske rabe in slovarskih uporabnikov, ki bi jo sodoben slovaropisni projekt moral – in glede na nove tehnološke možnosti mogel – zavzeti; opozoriti na nekatere uporabniške raziskave, ki so bile za slovenski prostor že pripravljene; in predstaviti potenciale, ki jih uporabniške raziskave ponujajo za nadaljnji razvoj področja.

2 KDO JE SLOVARSKI UPORABNIK?

Razpravo začenjamo z razmislekom, zakaj je (pre)splošno opredeljevanje ciljnega/potencialnega slovarskega uporabnika lahko strokovno nezadostno in v nadaljevanju

3 Primere je mogoče videti v Perdih (2009), vendar tega dela ne navajamo zato, ker bi se tovrstna praksa pojavljala samo pri skupini raziskovalcev, ki je na posvetu sodelovala, ampak zato, ker zbornik poleg strokovnih prispevkov vsebuje tudi zapis diskusije, v kateri so razmisleki o uporabnikih jasneje razvidni. Na podlagi teh zapisov P. Gantar denimo zaključi, da je »/s/plošni uporabnik, ki ga omenjajo udeleženci inštitutskega posveta kot dediščino SSKJ, /.../ v tem trenutku pravzaprav najmanj definiran uporabnik, še posebej, če v skladu s slovarsko tradicijo SSKJ govorimo o 'sodobnem človeku', saj se ta spreminja skladno s časom, v katerem živi.« (2014: 197—199).

predlagamo kot alternativno možnost nekoliko natančnejšo delitev slovarske rabe in uporabniških skupin, ki jih razumemo kot izhodišče za nadaljnje uporabniške raziskave, kot tudi določitev ciljnega uporabnika splošnega slovarja slovenskega jezika. Za izhodišče navajamo opredelitev, ki jo prinaša *Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika,* enega od dveh predlogov za izdelavo novega slovarja slovenščine, ki sta trenutno dostopna slovenski strokovni javnosti:[4]

> Ciljni uporabnik NSSKJ kot temeljnega sodobnega slovarja slovenskega knjižnega jezika je odrasli rojeni govorec slovenščine, ki bo predvidoma uporabljal zlasti elektronsko izdajo slovarja (tudi v povezavi z drugimi jezikovnimi priročniki, viri in zbirkami), obenem pa ne bo prikrajšan za klasično, tiskano obliko. Slovar bo sestavljen tako, da ga bodo lahko uporabljali tudi ostali kompetentni govorci slovenščine (Gliha Komac et al. 2015: 4).

Ciljni uporabnik slovarja je torej opredeljen s tremi značilnostmi: (I) je odrasel, (II) slovenščina je njegov prvi jezik, (III) uporablja predvsem elektronsko obliko slovarja, vendar želi tudi tiskano. Zadnja poved odstavka se po vsej verjetnosti navezuje na prvi dve navedeni točki: ob ustrezajoči kompetentnosti bo uporabnik lahko uporabljal slovar, čeprav (še) ni odrasel ali slovenščina ni njegov prvi jezik. Če pustimo ob strani nekatera odprta vprašanja (Kako razumeti odraslost uporabnika, v pravnem, biološkem, sociološkem smislu? Kakšno potrebo ima uporabnik elektronske izdaje slovarja po tiskani obliki? Katere kompetence so predpogoj za uporabo slovarja?), je mogoče gornji navedek razumeti kot opredelitev najširše možne množice, ki jo slovarski projekt glede na izbrano metodologijo lahko cilja. Jasno je, da je ob pripravah priročnikov, kot je splošni enojezični razlagalni slovar – še zlasti če gre za projekt na področju z manjšim številom govorcev, kjer je slovarska produkcija omejena – široka opredelitev potencialne rabe edina možna, saj mora tak slovar odgovarjati na širok spekter uporabniških jezikovnih vprašanj. Prav tako je jasno, da nacionalni slovarski projekt ne more biti odvisen od želja in idej posameznih (potencialnih) slovarskih uporabnikov, tudi zato, ker so slednje lahko nerealistične ali pa si vsebinsko nasprotujejo, kot kaže prispevek Arhar Holdt et al. (2015). Vendar pa je podatke o uporabnikih kljub temu potrebno zbrati, raziskati in ustrezno generalizirano upoštevati ter v tem smislu poskrbeti, da splošno zasnovana izhodiščna kategorija ne ostane notranje nestrukturirana, neraziskana in nedomišljena.

Pri določanju ciljnega uporabnika se opiranje na predvidene kompetence kaže kot izmuzljiv kriterij, zlasti v povezavi s stopnjo izobrazbe,[5] saj ni jasno, ali naj bi šlo pri tem za opredeljevanje uporabnikovih kognitivnih sposobnosti, njegove

4 Nekoliko bolj razdelano, vendar še vedno precej splošno opredelitev ciljnega uporabnika prinaša Predlog za izdelavo Slovarja sodobnega slovenskega jezika (Kreket al. 2013b: 20–21).

5 Po podatkih SURS je imelo 1. januarja 2014 med prebivalci nad 15 let – teh je bilo 1.760.032 – osnovnošolsko izobrazbo ali manj 26,7 %, srednješolsko 52,7 % in višjo 20,5 % prebivalcev (Podatki SURS 2015a). Pri tem je seveda treba upoštevati, da so kategorije notranje zelo raznolike (srednješolska izobrazba, o kateri se pogosto govori v povezavi z uporabniki splošnega razlagalnega slovarja, se razpenja od nižjih poklicnih šol do gimnazij) in da se izobrazbena struktura v starostnem prerezu razlikuje.

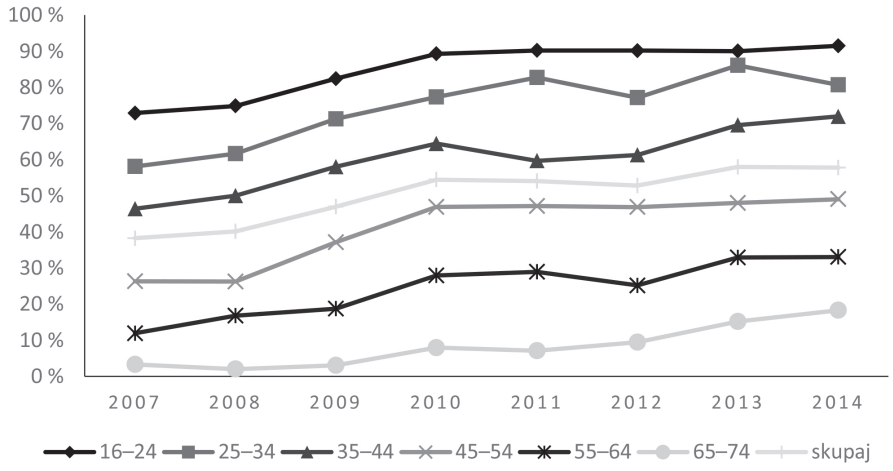
seznanjenosti z metajezikovnim aparatom, opismenjenostjo za delo s specifičnimi jezikovnimi priročniki ali česa četrtega. V primeru, da bi se za tovrstni kriterij vseeno odločili, je treba izhodišče natančneje definirati, razpravo pa podkrepiti s podatki o tem, če in kako določene skupine (obstoječih ali predvidenih) uporabnikov slovarske podatke de facto razumejo; presojanje o razumljivosti, intuitivnosti in podobnih značilnostih slovarja s strani slovaropiscev samih je namreč nezanesljivo, saj slednji praviloma spadajo v skupino najviše izobraženega prebivalstva, specializiranega za jezikoslovje, in so v tem smislu izjemno atipični slovarski uporabniki. V slovenskem prostoru so bile do sedaj opravljene raziskave o razumevanju podatkov iz SSKJ pri učencih (Rozman 2010; Rozman et al. 2010, pregled rezultatov ponuja Rozman et al. 2015). Za odraslo populacijo primerljivih podatkov še nimamo, je pa bilo ob upoštevanju izsledkov in dobrih praks iz evropskega prostora veliko strokovnega razmisleka posvečenega zahtevnosti definicijskega jezika SSKJ (npr. Kosem 2006; Gantar in Krek 2009).

Ob razpravi o kompetencah se odpira vprašanje skupin, ki jih uvodni navedek ne vključuje kot primarno ciljne, torej govorcev, ki jim slovenščina ni prvi jezik, in mlajših govorcev. Po podatkih SURS je imela 1. januarja 2015 Slovenija 2.062.874 prebivalcev, od tega 101.532 tujih državljanov oz. 4,9 % vseh prebivalcev Slovenije (Poročilo SURS 2015a), kar pomeni, da ta skupina nikakor ni zanemarljiva, še zlasti, če k njej prištejemo govorce slovenščine, ki bivajo izven meja Republike Slovenije (glej razdelek 3.4). K temu lahko dodamo 75.325 mladih (Podatki SURS 2015b), ki so bili v začetku šolskega leta 2014 vključeni v srednješolsko oz. 169.101 (Podatki SURS 2015c) vključenih v osnovnošolsko izobraževanje. Ali pripravljavci slovaropisnega projekta naštete skupine (oz. njihove določene segmente) v izhodišču upoštevajo ali ne, je odvisno od namena projekta, zmotno pa bi bilo računati s tem, da se lahko za te uporabnike post festum prilagajajo leksikalni podatki, ki niso že od začetka upoštevali njihovih specifik. Z drugimi besedami, če se naštete uporabniške skupine uvrstijo med ciljne naslovnike slovarja, je treba njihove specifike raziskati in upoštevati od samega začetka snovanja slovarske baze naprej (gl. tudi prispevek Rozman et al. 2015).

Čeprav o primarnosti digitalne oblike slovarskih podatkov ni več dvoma, je nekaj statistik mogoče navesti tudi glede tehnološke opremljenosti prebivalstva. Mobilne storitve so še vedno v porastu: v primerjavi s predhodnim letom je v letu 2014 število uporabnikov mobilnega omrežja zraslo za 2 % (in ob koncu leta 2014 preseglo številko 2.326.000), skupni odhodni promet iz mobilnih omrežij se je povečal za 4 %, poslanih je bilo za 13 % več SMS-ov in za 11 % več MMS-ov. Raste tudi dostopnost interneta: od leta 2013 do 2014 se je število širokopasovnih internetnih priključkov povečalo za 5 % oz. na 556.000 enot, od katerih jih je bilo 86 % v gospodinjstvih (Poročilo SURS 2015b). Dostopnost do interneta (ne glede na vrsto povezave) v slovenskih gospodinjstvih je v letu 2014 doseglo 77 %

(Podatki SURS 2015d), glavni razlogi za nedostop pa so previsoki stroški opreme (pri 11 % vprašanih) ali dostopa (pri 10 % vprašanih),[6] in ocena, da interneta ne potrebujejo (pri 16 % vprašanih) ali imajo o njem pomanjkljivo znanje (pri 14 % vprašanih) (Podatki SURS 2015e). Še prehod od gospodinjstev k posameznikom: v letu 2014 je uporabljalo internet skoraj vsak dan 58 % oseb, 42 % oseb je internet uporabljalo prek mobilnega telefona, prenosnega ali tabličnega računalnika ali drugih mobilnih naprav zunaj doma ali delovnega mesta. E-pošto je v letu 2014 pošiljalo ali prejemalo 62 % oseb, v spletnih družabnih omrežjih je sodelovalo 42 % oseb, storitve računalništva v oblaku pa jih je uporabljalo 31 % (Poročilo SURS 2015c). Te številke so nekoliko nižje od povprečja EU-28, kar nakazuje, da se bo rast v prihodnje še nadaljevala.

Navedeni podatki nakazujejo razvojne trende, ki jih morajo slovarski projekti upoštevati, vendar se je treba tudi pri tem izogibati pretiranim posplošitvam, kar ponazarja Graf 2. V grafu vidimo delež prebivalstva, ki internet uporablja (skoraj) vsak dan, pri čemer so podatki prikazani po starostnih skupinah (Podatki SURS 2012f). Že ta povsem osnovna delitev kategorije »odraslega potencialnega slovarskega uporabnika« razkrije bistvene razlike glede možnosti in predvidenih prioritet slovarske rabe, povprečje pa se v sliki zarezuje precej varljivo.



Graf 1: Delež prebivalstva, ki uporablja internet vsak dan ali skoraj vsak dan, glede na starost.

Če zaključimo gornjo razpravo: pri opredeljevanju ciljnega oz. potencialnega uporabnika kateregakoli slovarja je treba poiskati srednjo pot med pomensko praznimi

6 Ta podatek je za razpravo koristen zato, ker nakazuje, da vprašanja dostopnosti slovarja ne gre povezovati le z medijem oz. obliko, ampak tudi s finančno dostopnostjo: od uporabnikov, ki si ne morejo privoščiti dostopa do interneta, je najbrž težko pričakovati, da bodo veliko denarja namenili za slovarje, ne glede na njihovo obliko.

generalizacijami in množico posameznih uporabniških mnenj. Pri iskanju te poti se je treba opirati na podatke o uporabnikih, v primeru, da ti podatki niso na voljo ali niso dovolj natančni, pa pripraviti načrt za njihovo pridobitev. Ob tem se zdi nujno razpravo obrniti od uporabniških kompetenc k uporabniškim potrebam in s tem obrniti celotno logiko odnosa do uporabnika: ne zgolj identificirati uporabnika, ki bo določeno vrsto slovarja *zmogel* uporabljati, ampak opredeliti uporabnika, ki ga želimo s slovarskimi podatki opremiti, in z ustreznimi prilagoditvami metodologije izdelave slovarja omogočiti, da bo cilj v karseda visoki meri dosežen. Samo v tem primeru sploh lahko govorimo o *ciljnem* uporabniku, saj *ciljnost* implicira pripravljenost na prilagoditve naslovniku – kar brez podatkov o dejanski slovarski rabi in dejanskih slovarskih uporabnikih ni mogoče, saj »nima smisla govoriti o uporabniških potrebah, če na slednje gledamo abstraktno, ne da bi jih povezali s specifičnimi tipi uporabnikov in situacij« (Tarp 2009: 279, prevod Š. A. H.).

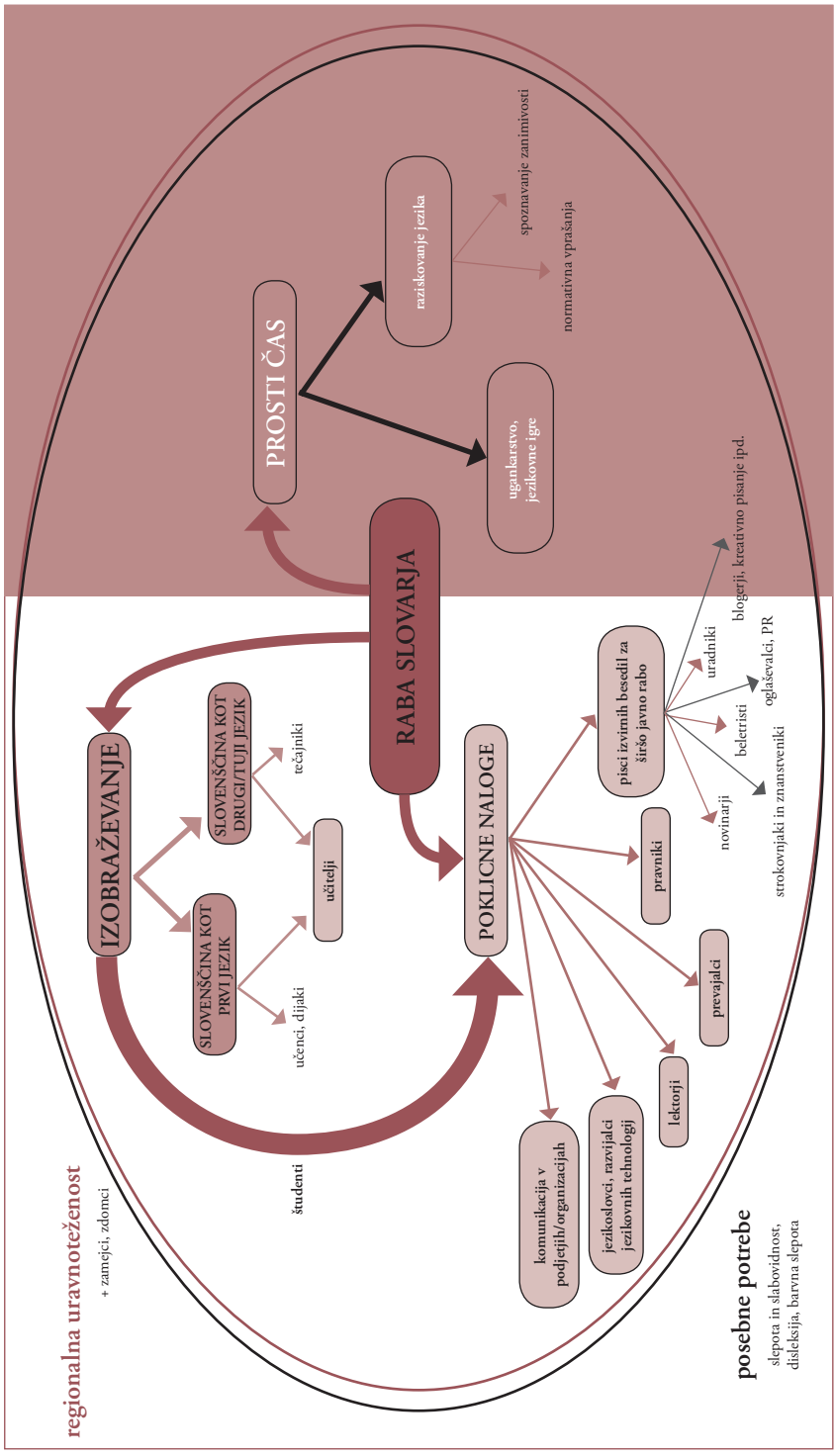
3 SITUACIJE SLOVARSKE RABE IN UPORABNIŠKE SKUPINE

Kot možno rešitev dileme v slovaropisnem odnosu do ciljnega/potencialnega uporabnika v tem poglavju predstavljamo shemo situacij slovarske rabe oz. uporabniških skupin. Situacije, v katerih se pogosto uporablja enojezični razlagalni slovar, smo razdelili v tri skupine: (I) raba slovarja v procesu (formalnega) izobraževanja, (II) raba slovarja za poklicne namene in (III) raba slovarja v prostem času, s čimer sledimo delitvi, ki jo v raziskavi navad uporabnikov danskega *Den Danske Ordbog* uporabljata Lorentzen in Theilgaard (2012).[7] Vsako od naštetih skupin smo nadalje členili in posamezne skupine povezali, kot prikazuje Slika 1. Pomembno je upoštevati, da shema ni razumljena kot zaključena ali dokončna, niti ne kot edina možna rešitev, temveč zgolj kot potencialno izhodišče za nadaljnje empirične raziskave.

3.1 Slovar v procesu izobraževanja

Ob snovanju slovaropisnih projektov je pomembno imeti pred očmi vlogo, ki jo slovar igra v procesu izobraževanja, z upoštevanjem razlik glede (slovarskih in jezikovnih) potreb in zmožnosti pri različnih uporabniških skupinah, ki jih kategorija pokriva. Splošno rečeno mora slovar v izobraževanju na eni strani odgovarjati jezikovnodidaktičnim željam in potrebam učiteljev, na drugi strani pa kognitivnim zmožnostim učencev in dijakov, ki se na različnih stopnjah izobraževalnega procesa razlikujejo. Ob tem je potrebno ločevati med učečimi, ki se s slovenščino

7 Rezultati njune raziskave kažejo, da redni uporabniki v splošnem slovar najpogosteje uporabljajo v službi (37 %), sledi raba v prostem času oz. doma (30 %) in v procesu izobraževanja (28 %). Navajata tudi kategorijo Drugo (5 %), ki ni razložena, mogoče pa je predvidevati, da gre za primere na mejah predlaganih kategorij (Lorentzen in Theilgaard 2012: 685).



Slika 1: Slovarska raba in uporabniške skupine.

srečujejo kot prvim jezikom, in tistimi, ki jim je slovenščina drugi ali tuji jezik, in upoštevati, da je tudi slednja skupina notranje zelo heterogena. (Faktor, ali je slovenščina za uporabnika prvi ali drugi/tuji jezik, sicer ni vezan le na proces izobraževanja, vendar pa je specifike teh uporabniških skupin v procesu izobraževanja najlažje preučevati.) Slovarju v procesu izobraževanja se posveča prispevek Rozman et al. (2015), ki prinaša izčrpen pregled obstoječih slovenskih in tujih raziskav s tega področja, v diskusiji pa opredeljuje glavne izzive in vizijo priprave enojezičnega razlagalnega slovarja na način, da bo uporaben tudi za šolsko rabo.

Kot posebna uporabniška skupina so na Sliki 1 na prehodu med izobraževanjem in poklicno rabo izpostavljeni študenti različnih študijskih smeri. To skupino je smiselno izpostaviti (tudi) zato, ker izkušnje iz tujine kažejo, da so študentje zaradi statusa in dostopnosti raziskovalnim projektom pogosto vključeni v uporabniška testiranja in evalvacije. Ob vključevanju študentov v uporabniške raziskave je treba upoštevati, da njihove navade, želje in preference niso nujno reprezentativne za celotno populacijo (Tarp 2009: 290), pri določenih temah (npr. raziskavah rabe slovarjev ali pridobivanja jezikovne kompetence v času študija) pa je ta skupina smiselna in smotrna izbira.

3.2 Slovar za poklicno rabo

Enako kot v izobraževanju so uporabniške potrebe raznolike, če jih preučujemo v kontekstu poklicne rabe. Morda prav ta najbolje izpričuje neustreznost koncepta splošnega uporabnika, saj je že na prvi pogled jasno, da se raba slovarja razlikuje, če gre za potrebe prevajanja, lektoriranja, pisanja strokovnega članka ali razlago uradnega dopisa. Slika 1 navaja osnovne uporabniške skupine pri poklicni rabi, pri čemer je potrebno ponoviti, da gre le za izhodiščno strukturo in notranje heterogene skupine, ki jih je v nadaljevanju treba natančneje raziskati.

Prva gruča združuje uporabniške skupine, ki smo poimenovali: novinarji, uradniki, strokovnjaki oz. znanstveniki, ustvarjalci leposlovja, oglaševalci in blogerji, torej glede na poklice oz. aktivnosti, pri katerih se pojavlja (redna) produkcija besedil, namenjenih širši javni rabi. Tem skupinam se posveča Mikolič (2015), ki na osnovi intervjujev med predstavniki znanstvenikov, literatov, oglaševalcev in novinarjev ugotavlja, da se glede rabe slovarja med posameznimi skupinami kažejo stične točke, vendar pa vsaka od njih prinaša tudi specifike, ki jih je brez podatkov težko predvideti, jih je pa pri snovanju slovarja treba upoštevati. Naslednja gruča je sestavljena iz uporabniških skupin, ki tvorijo različne vrste besedil v poslovni komunikaciji, npr. vodstveni, računovodski kadri, poslovna asistenca, uporabniška podpora, komunikacije z javnostmi. Ti profili so glede rabe jezikovnih priročnikov slabo raziskani, čeprav je mogoče sklepati, da jezikovni problemi,

o katerih uporabniki poročajo po jezikovnih forumih in svetovalnicah, v določeni meri izvirajo iz teh uporabniških skupin (več o tem Arhar Holdt et al. 2015). Sledijo še druge uporabniške skupine, ki jih je potrebno obravnavati ločeno: lektorji, pri katerih se predvidevajo specifike rabe normativnih podatkov; jezikoslovci in drugi raziskovalci, pri katerih se predvideva uporaba slovarskih podatkov za raziskovalno/razvojno gradivo;[8] pravniki, pri katerih se predvideva specifične načine rabe slovarskih podatkov za tvorbo in interpretacijo pravnih besedil; in nenazadnje prevajalci, raziskavo katerih predstavljajo Čibej et al. (2015) kot primer profiliranja potreb določene uporabniške skupine.

3.3 Slovar za prostočasne aktivnosti

Ko govorimo o rabi slovarja v prostem času, ne gre za opredelitev, kdaj v dnevu se slovar uporablja, ampak za namen rabe, ki v tem primeru ni neposredno vezan na proces strukturiranega izobraževanja ali specifične poklicne naloge. Prva in morda najbolj specifična skupina v tej kategoriji so uporabniki, ki slovarje uporabljajo kot referenčne priročnike za ugankarske namene ali igranje jezikovnih iger, kot je npr. križemkražem (igra, podobna angleški *Scrabble*).[9] V drugi skupini so uporabniki, ki jih jezik zanima bodisi v normativnem smislu (ne/ustreznost določene jezikovne rabe), bodisi v smislu raziskovanja zanimivosti (bogatenje besedišča, spoznavanje frazeologije, etimologije). V zadnji skupini so uporabniki, ki slovar uporabljajo zgolj sporadično, vendar imajo vseeno določena pričakovanja glede tega, kakšno vlogo naj bi ta priročnik v družbi imel in katere vsebine naj bi prinašal. Prvi slovenski vpogled v potrebe in mnenja naštetih uporabniških skupin prinaša prispevek (Arhar Holdt et al. 2015).

3.4 Slovar za različne jezikovne skupnosti

Splošni enojezični razlagalni slovar mora pri uporabniških raziskavah zagotoviti ustrezno regionalno uravnoteženost in ob tem vključiti tudi govorce slovenščine, ki bivajo izven meja Republike Slovenije (zamejce in zdomce). Na drugi strani je treba upoštevati uporabnike, ki imajo zaradi različnih oviranosti specifične sporazumevalne potrebe. O upoštevanju različnih skupin, ki spadajo v to kategorijo,

8 V tem prispevku se osredotočamo na slovar kot jezikovni pripomoček za človeškega uporabnika, ob strani pa puščamo (za sodobno slovaropisje sicer ključna) vprašanja priprave in dostopnosti jezikovnih podatkov za strojno rabo oz. razvoj jezikovnih tehnologij. Ko na tem mestu govorimo o potrebah jezikoslovcev in drugih k jeziku usmerjenih raziskovalcev, imamo torej v mislih predvsem potrebe, ki so rešljive v slovarskem vmesniku (npr. možnosti naprednega iskanja, sintetičnega prikazovanja podatkov, izvažanja podatkov ipd.). Vprašanja, kako pripraviti slovarske podatke za razvoj jezikovnih tehnologij, presegajo metodologijo področja uporabniških raziskav, zato se jim je treba posvetiti ločeno. Več o snovanju slovarske baze za človeškega ter strojnega »uporabnika« je mogoče prebrati npr. v Gantar (2015).

9 Povezava na stran društva Križemkražem s tekmovalnim pravilnikom, ki opredeljuje referenčne priročnike je na: http://www.drustvo-krizemkrazem.si/ (dostop 8. 8. 2015).

na deklarativni ravni velja visok družbeni konsenz, vendar se v praksi pri pripravi jezikovnih priročnikov te specifike zelo malo upošteva. Specifične raziskave o prilagoditvah jezikovnih priročnikov uporabnikom z okvarami vida ali boleznimi oči v slovenskem prostoru še ne obstajajo, velja pa načelna smernica, da je digitalna oblika slovarja zanje bistveno prijaznejša od knjižne, saj omogoča prilagoditve, ki na papirju niso mogoče.[10] Nekatere od teh prilagoditev so za izvedbo zelo preproste, npr. izbira fontov, ki lajšajo branje osebam z disleksijo, ali možnost povečave črk znotraj obstoječega formata.

Kot je bilo napisano v uvodu v razdelek 3, lahko naštete uporabniške skupine razumemo kot izhodišče za nadaljnje raziskave uporabniških navad, potreb, želja in mnenj, pri čemer se je mogoče poslužiti uveljavljene metodologije področja ali razmisliti o inovativnih pristopih (gl. npr. Tarp 2009). Različne metodološke možnosti preizkušajo prispevki, ki se s slovarskimi uporabniki ukvarjajo v nadaljevanju te monografije, razpravo v tem članku pa nadaljujemo s predstavitvijo drugega pomembnega potenciala, ki ga slovaropisju s svojo odprtostjo, pretočnostjo in razširjenostjo ponuja digitalni medij: vključevanju (potencialnih) slovarskih uporabnikov v sam postopek priprave slovarja.

4 VKLJUČEVANJE UPORABNIKOV V PRIPRAVO SLOVARJA

Sodobni slovaropisni projekti se različnih možnosti vključevanja (potencialnih) slovarskih uporabnikov poslužujejo v različni meri, odvisno od tega, v kakšnem okolju in za kakšen namen določen slovar nastaja. V tem prispevku možnosti vključevanja uporabnikov delimo na tri skupine: (I) kot evalvatorje predvidenih slovarskih rešitev, (II) kot sodelavce pri pripravi jezikovnih podatkov in (III) kot zainteresirano uporabniško skupnost, ki s komentarji objavljenih slovarskih podatkov vpliva na razvoj vira. Naštete točke se, kot rečeno, navezujejo na sodelovanje z uporabniki med samim slovaropisnim procesom, neodvisno od slednjega pa lahko potekajo druge raziskovalne aktivnosti, npr. prej omenjene raziskave uporabniških skupin.

S prehodom v digitalni medij postaja slovar prilagodljiv na načine, ki v preteklosti niso bili mogoči,[11] nove možnosti pa pripravljavce slovarja postavljajo pred nove odločitve. Med pripravo slovarja, ki se želi glede vsebine, oblike in funkcionalnosti

10 Korespondenca z Matejo Forte, ki v sklopu doktorske raziskave pripravlja smernice za pripravo (primarno učbeniškega) gradiva slepim in slabovidnim.

11 Nekaj primerov slovaropisnih premikov v smer uporabniške prijaznosti in prilagodljivosti slovarja našteva npr. de Schryver (2009: 182–184), pri čemer prilagodljivosti slovarja v tem prispevku ne razumemo kot (enkratno in dokončno) prilagoditev posameznemu uporabniku oz. uporabniški skupini, ampak v smislu dinamičnega prikaza slovarskih informacij glede na različne vrste poizvedb, ki jih je mogoče ustrezno razumeti zgolj z raziskovanjem slovarske rabe in jezikovnih dilem govorcev in govork. O tem Arhar Holdt et al. (2015).

kar najbolj približati ciljnemu uporabniku, je zato smiselno uporabniške preference preverjati sproti med samim procesom odločanja. Digitalno okolje omogoča dolgotrajnejše sodelovanje s stalnim (ustrezno vzorčenim) naborom uporabnikov, ki na pobudo strokovnjakov odgovarjajo na vprašanja in izvajajo krajše evalvacije, ob čemer se odgovori programsko beležijo in statistično obdelujejo. Pri načrtovanju tovrstnega sodelovanja seveda ne gre pričakovati, da bodo uporabniki samoiniciativno razvijali predloge za konceptualne nadgradnje slovarja, lahko pa izberejo, katera od dveh predlaganih možnosti se jim zdi preglednejša ali uporabnejša za njihovo delo. Predlog za izvedbo tovrstnih testiranj predstavlja prispevek Fišer et al. (2015). Poleg tovrstnih sprotnih ocen lahko uporabniki v kasnejših fazah projekta izvedejo tudi celovitejše evalvacije, npr. evalvacije uporabniškega vmesnika, pri katerih je mogoče uporabiti tudi katero od naprednejših metod opazovanja slovarske rabe (gl. razdelek 1).

Druga možnost sodelovanja je v smislu izrabe uporabniške pomoči za selekcioniranje, urejanje in dopolnjevanje jezikovnih podatkov po principu množičenja. Tovrstno sodelovanje lahko pomembno izboljša časovno in finančno sliko projekta, ob čemer je samoumevno, da morajo biti naloge za množičenje izbrane premišljeno, da ne vplivajo na zanesljivost slovarskih informacij. O množičenju v slovaropisju v tej monografiji pišeta Fišer in Čibej (2015).

Nenazadnje pa je uporabnike slovarja mogoče razumeti kot spletno skupnost in jim v okviru nastajajočega slovarja ponuditi prostor za izmenjavo mnenj oz. sporočanje komentarjev in vprašanj, pri čemer je poleg interakcije med uporabniki samimi mogoče vzpostaviti tudi povezavo med uporabniki in pripravljavci slovarja. Ob tem je pomemben podatek, da se možnosti sodelovanja v digitalnem okolju zavedajo tudi (že) uporabniki sami, ki si vključevanja želijo in ga v določeni meri celo pričakujejo (Müller-Spitzer 2014: 156–159). Obstoječi digitalni slovarji se trenutno poslužujejo predvsem vključevanja forumov in klepetalnic ter povezave z družbenimi omrežji, torej izvedbeno relativno preprostih rešitev. V raziskovalnem smislu pa je vsako uporabniško akcijo ter interakcijo mogoče videti tudi kot potencial za sistematično zbiranje podatkov, ki lahko vodijo v nadaljnji razvoj slovarske vsebine in oblike.

5 SKLEP IN NALOGE ZA NAPREJ

V prispevku smo argumentirali, da je koncept *splošnega* (kot tudi *povprečnega* ali *zahtevnega*) uporabnika slovarja – ali katerega koli drugega jezikovnega priročnika – za strokovno razpravo brezpredmeten oz. nezadosten, če mu ne sledijo natančnejše opredelitve (razdelek 2).[12] Pri tem ne zanikamo dejstva, da mora biti

12 Problematičen je tudi pojem *splošne rabe* slovarja, čemur se prispevek sicer ne posveča eksplicitno, vendar je težavo mogoče razumeti v primerljivem smislu.

priprava splošnega enojezičnega razlagalnega slovarja usmerjena k najširši možni rabi, vendar pa široke rabe ni dovolj le predvidevati, pač pa jo je treba razisk(ov)ati, identificirati potrebe, želje in zmožnosti specifičnih uporabniških skupin in izsledke (kot tudi uporabnike same) ustrezno vključiti v slovaropisni proces. Razpravo zato sklepamo z opredelitvijo glavnih prioritet za nadaljnji razvoj področja uporabniških raziskav, ki bi jih bilo smiselno vključiti v načrte za pripravo novega enojezičnega razlagalnega slovarja za slovenščino:

A - Analizirati in v slovaropisnem kontekstu osmisliti tiste podatke o uporabnikih, ki so že na voljo, npr. podatke o jezikovnih dilemah govork in govorcev slovenščine in podatke o rabi obstoječih jezikovnih priročnikov (npr. dnevniške datoteke iskanj po slovarjih, leksikonih besednih oblik, korpusih ipd.).

B - Sistematično raziskati uporabniške navade, potrebe in želje glede slovarske rabe, pri čemer je mogoče izhajati iz (po potrebi nadgrajene) sheme uporabniških skupin (razdelek 3). Za tovrstne raziskave bi bilo smotrno uporabiti obstoječo metodologijo področja (razdelek 1), npr. ankete, intervjuje in testiranja, vendar z upoštevanjem opozoril stroke (npr. Lew 2003; Tarp 2009; Müller-Spitzer 2014).

C - Zasnovati slovarski projekt na način, da bo omogočeno sodelovanje s (potencialnimi) slovarskimi uporabniki oz. njihovo vključevanje, pri čemer je vsakršno uporabniško aktivnost treba razumeti kot dragoceno informacijo za nadaljnji razvoj slovarja (razdelek 4).

Ob vseh naštetih točkah je pomembno v slovaropisnem diskurzu vztrajati pri ločevanju *védenja* o uporabnikih, ki izvira iz podatkov in raziskav, ter *predvidevanj* o uporabnikih, ki predstavljajo legitimno izhodišče za razpravo, vendar v metodološkem smislu ne morejo in ne smejo zadoščati. Prepričanja slovaropiscev o uporabniških zmožnosti in potrebah so namreč – po duhovitih besedah G. Hatheralla – primerljiva predstavam angleške princese Anne, kako je živeti ob denarni socialni pomoči: v grobem so lahko kar blizu resnice, ko pride do podrobnosti, pa še zdaleč ne (Hatherall 1984: 183).

Slovarji in učenje slovenščine

Tadeja Rozman, Iztok Kosem, Nataša Pirih Svetina in Ina Ferbežar

Abstract
This paper discusses which Slovenian dictionary or dictionaries would be the most suitable for native Slovenian speakers and non-native Slovenian speakers to use. Slovenian studies are presented that focus on dictionary use and the comprehensibility of dictionary information among Slovenian elementary and secondary school students as well as non-native Slovenian speakers. A brief overview is also presented of relevant findings from dictionary use studies conducted abroad. After this overview of the needs, abilities and preferences of dictionary users who are learning a language, the paper concludes with some suggestions for Slovenian dictionary makers.

Keywords: dictionary use, school dictionary, learner's dictionary, vocabulary acquisition, language learning

Ključne besede: raba slovarjev, šolski slovar, slovar za tujce, usvajanje besedišča, učenje jezika

1 UVOD

Slovarji sodijo med osnovne jezikovne priročnike, zato so nepogrešljivi pri učenju tujih jezikov, pozitivno vlogo pa igrajo tudi v procesu usvajanja prvega jezika. Uporaba slovarjev namreč pozitivno vpliva na uspešno učenje novih besed ter pripomore k poglabljanju znanja o besedah tako na ravni pomena kot rabe (Paynter et al. 2005: 35–37, 41–45), bogat besedni zaklad pa je izjemno pomemben element sporazumevalne zmožnosti posameznikov – v kontekstu izobraževanja velja poudariti, da med drugim igra veliko vlogo pri šolskem uspehu, saj omogoča lažje razumevanje novih šolskih vsebin ter ima pozitiven vpliv na bralno pismenost (Paynter et al. 2005: 3–7, Pečjak 2012: 31). Vendar pa strokovnjaki opozarjajo (npr. Wright 1998: 7), da ima uporaba slovarja, ki ne ustreza spoznavno-jezikovnemu razvoju otrok in mladostnikov in ne upošteva specifičnih potreb ciljnih uporabnikov, lahko ravno nasprotne, torej precej negativne učinke.

V Sloveniji je razmislekov o oblikovanju slovarjev, ki bili primerni za tuje uporabnike ali učence v procesu osnovnošolskega in srednješolskega izobraževanja, razmeroma malo, prav tako ne obstajajo slovarji, narejeni prav za ti dve ciljni skupini. Zato učitelji slovenščine (kot prvega in kot drugega in tujega jezika) pri pouku (pogosto) uporabljajo kar splošni slovar, torej Slovar slovenskega knjižnega jezika (SSKJ),[1] in to kljub temu da je v osnovi namenjen odraslim domačim govorcem slovenščine. Problematičnosti uporabe oz. razmeroma omejene uporabnosti tega slovarja pa se, kot kažejo raziskave,[2] zaveda le manjši del učiteljstva, kar še posebej velja za učitelje v osnovnih in srednjih šolah,[3] ki so večinoma mnenja, da je za šolsko rabo SSKJ povsem primeren. Deloma je to seveda posledica odsotnosti drugih in drugačnih slovarskih virov, prevladujoče simbolne vloge splošnega slovarja, ki mu družba priznava status temeljnega jezikovnega priročnika (ker je v njem zajeto besedno bogastvo slovenščine, zato predstavlja temeljni vir za črpanje znanja o slovenskem besedišču, prim. Stabej 2009, Rozman 2009), deloma pa tudi zaradi pomanjkanja strokovnih objav in raziskav, ki bi se v našem prostoru ukvarjale z didaktičnimi vidiki uporabe slovarja ter s procesi usvajanja in učenja besedišča.

V zadnjem desetletju so se sicer zgodili nekateri pozitivni premiki na tem področju, saj je bilo tudi pri nas opravljenih nekaj raziskav na temo uporabe in razumljivosti slovarjev v okviru izobraževanja. Rezultati takšnih raziskav so za pripravo slovarskih virov izjemno pomembni. Izhodišče sodobnih leksikografskih praks namreč je, da je pri pripravi slovarjev potrebno upoštevati potrebe in zmožnosti ciljne publike, deloma pa tudi navade glede iskanja informacij o

1 O tem podrobneje v nadaljevanju članka v razdelku 2.
2 Gl. razdelek 2.
3 Manj pa za učitelje slovenščine kot drugega in tujega jezika.

jeziku – pri snovanju slovarjev za šolsko rabo oz. tuje govorce je torej poleg poznavanja splošnih značilnosti spoznavno-jezikovnega razvoja in specifičnosti usvajanja besedišča potrebno tudi védenje o tem, katere podatke o besedišču bodo uporabniki na določenih stopnjah izobraževanja oz. stopnjah jezikovnega znanja verjetneje bolj potrebovali (in bodo zato za njih relevantni) ter kako naj bodo ti podatki zapisani, da bodo razumljivi (in zato uporabni). Seveda se do določene mere lahko naslonimo na izkušnje iz tujine – v mednarodnem prostoru je bilo opravljenih že precej raziskav o uporabniških vidikih slovarjev,[4] prav tako imajo izkušnje z izdelavo različnih šolskih slovarjev in slovarjev za tujce – vendar pa tuje prakse ne morejo biti v slovenski prostor prenesene brez upoštevanja specifičnosti tako slovenskega jezika kot prostora, s čimer imamo v mislih predvsem družbena stališča do jezikovnih priročnikov, standardizacije in poučevanja jezika, pa tudi didaktična načela in prakse slovenskega izobraževalnega sistema.

Zaradi omejenosti s prostorom se bomo v prvem delu prispevka osredotočili na prikaz raziskav, opravljenih v Sloveniji (razdelek 2), ga dopolnili s spoznanji iz tujine (razdelek 3), nato pa v razdelku 4 razmišljali o konkretnih rešitvah, ki jih lahko izpeljemo na podlagi opravljenih raziskav in poznavanja področja, kar zajema tudi upoštevanje značilnosti spoznavno-jezikovnega razvoja, ki ga v prispevku sicer podrobneje ne predstavljamo.

2 UPORABNIŠKE RAZISKAVE V SLOVENIJI

Pri nas je bilo v zadnjih letih opravljenih nekaj raziskav, ki se osredotočajo na razumljivost slovarskih podatkov med učenci, rabo slovarjev pri pouku slovenščine, deloma pa tudi na analizo leksikalnih težav. V tem poglavju bomo na kratko predstavili glavne rezultate, ki so relevantni za razmislek o pripravi slovarskih vsebin, primernih za ti dve ciljni skupini.

2.1

Prva obsežnejša anketna raziskava je bila narejena leta 2008 (Stabej et al. 2008). V raziskavo je bilo zajetih 409 učiteljev slovenščine in 3427 učencev od 4. razreda osnovne šole do 4. letnika srednje šole. Vsebinsko je bila anketa dvodelna: prvi del je preverjal uporabo in vrednotenje enojezičnih slovarskih priročnikov, drugi pa zaznavanje problemov pri usvajanju besedišča.

Odgovori so pokazali, da učitelji slovarje pri pripravi na pouk uporabljajo razmeroma pogosto, in sicer za različne vsebine: najpogosteje za obravnavo besedišča,

4 Gl. razdelek 3.

a tudi slovnice, umetnostnih in neumetnostnih besedil ter pri pripravi in popravljanju šolskih nalog in testov. Pri pripravi na pouk najpogosteje uporabljajo SSKJ, enako velja za delo v razredu – kar 96,8 % učiteljev je zatrdilo, da ga pri pouku uporabljajo vsaj občasno. SSKJ vključujejo pri obravnavi različnih učnih vsebin (Tabela 1), tudi ko delo s slovarjem ni predvideno po učnem načrtu ali v učbeniku, poleg tega pa pripravljajo tudi posebne vaje za spoznavanje SSKJ.

Tabela 1: Vključevanje SSKJ pri posameznih učnih vsebinah v razredu (% predstavljajo deleže učiteljev, ki uporabljajo SSKJ pri navedenih dejavnostih)

Dejavnost	%
obravnava neumetnostnega besedila	68,5
obravnava besedoslovja in frazeologije	65,3
obravnava umetnostnega besedila	56,2
obravnava pravopisa	52,8
skupinska poprava testov, domačih nalog	39,4
obravnava pravorečja	37,9
obravnava slovnice	32,3
obravnava besediloslovja in sporočanje	24,2
drugo	2,4

Velika večina učiteljev je tudi zatrdila, da spodbujajo učence k samostojni rabi slovarjev pri različnih, predvsem produktivnih sporazumevalnih dejavnostih (Tabela 2), v primeru nerazumevanja besed pa učencem vsaj občasno naročijo, naj pomen poiščejo v slovarju.

Tabela 2: Spodbujanje rabe slovarjev[5] pri posameznih dejavnostih (% predstavljajo deleže učiteljev, ki spodbujajo rabo slovarjev pri navedenih dejavnostih)

Dejavnost	%
pisanje besedil	71,6
iskanje sopomenk in protipomenk	62,3
priprava govornih vaj	61,1
poprave besedil	56,0
iskanje domačih ustreznic	55,5
iskanje slogovno nezaznamovanih ustreznic	54,0
reševanje jezikovnih vaj	37,2
branje	27,6
drugo	4,9

5 Vprašanje se je nanašalo na rabo SSKJ, slovarskega dela Slovenskega pravopisa in slovarjev tujk.

Učitelji se strinjajo, da je spoznavanje enojezičnih slovarjev v šoli koristno, ker pripomore k boljši sporazumevalni zmožnosti učencev, jim pomaga pri pravilnejši rabi jezika, igra pomembno vlogo pri usvajanju besedišča in širjenju besednega zaklada in ker popisuje besedno bogastvo slovenščine. O SSKJ na sploh imajo pozitivno mnenje:

- menijo, da je koristen pripomoček pri reševanju različnih jezikovnih težav (predvsem pri razumevanju besed ter reševanju težav z zapisom in izgovorom besed, nekoliko manj tudi pri vprašanjih sloga, terminologije, pragmatike in slovnice besed),
- se strinjajo z njegovo normativno veljavnostjo,
- menijo, da je razumljiv in da ga ni težko uporabljati,
- velika večina (73,3 %) pa je tudi mnenja, da je primeren za učence.

Podobno kot učitelji imajo tudi učenci o slovarjih dobro mnenje:

- menijo, da pomaga pri pravilni rabi slovenščine in pri reševanju težav s slovenščino,
- slovarje dojemajo kot normativne priročnike,
- razlage v slovarjih se jim ne zdijo pretežke, se pa strinjajo s trditvijo, da je slovarje težko uporabljati zaradi slabo razumljivih krajšav in znakov.

Vendar pa kljub pozitivnemu mnenju velika večina učencev ne mara uporabljati slovarjev (s trditvijo »Rad/a uporabljam slovarje« se je strinjalo le 37 % osnovnošolcev in 30,3 % srednješolcev) niti jih ne uporablja (SSKJ npr. uporablja le 24,5 % osnovnošolcev in 16,5 % srednješolcev). Na vprašanje o samostojni rabi slovarjev doma so učenci odgovorili, da slovarje uporabljajo pretežno takrat, ko morajo rešiti naloge, povezane s slovarji, pri vprašanju, ki je preverjalo rabo slovarjev pri posameznih težavah z besedami, pa se je izkazalo, da jih najpogosteje uporabijo kot pomoč za razumevanje in zapis besed (čeprav so deleži učencev, ki slovarje uporabljajo, razmeroma nizki). Prav tako se je izkazalo, da je uporaba slovarja pri reševanju leksikalnih težav ena izmed najmanj priljubljenih strategij (raje vprašajo, se nalogi izognejo ali pogledajo na internet).

Na tem mestu so zanimivi tudi odgovori, katere vrste napak v rabi besedišča učitelji najpogosteje opažajo pri pisanju in govoru učencev: daleč najpogostejše so pravopisne oz. izgovorne napake, sledijo jim slovnične in slogovne, manj pogoste pa so pomenske, kolokacijske, sintagmatske in frazeološke napake.

2.2

Podobna, vendar bistveno manj obsežna anketna raziskava je bila opravljena leta 2013 (Čebulj 2013), v katero je bilo zajetih 75 učiteljev razrednega pouka. Tudi

tukaj se je izkazalo, da je večina že uporabila SSKJ pri pouku (celo v 1. razredu) in da učence učijo uporabljati SSKJ, slovarje pa uporabljajo tudi kot eno izmed strategij za razlaganje pomena besed.[6] Učitelji sicer zaznavajo nekaj težav, ki jih imajo učenci z uporabo SSKJ (največ težav opažajo pri iskanju gesel po abecedi) in se večinoma strinjajo, da bi potrebovali enojezični slovar, namenjen šolarjem.

2.3

V okviru projekta Sporazumevanje v slovenskem jeziku (SSJ)[7] je bila leta 2010 izvedena zelo obsežna anketna raziskava o jezikovnem pouku slovenščine (Rozman et al. 2010; 2012). Vanjo je bilo vključenih 276 učiteljev slovenščine ter 1465 učencev 3. triletja osnovnih šol in srednješolcev. Vprašanj, ki bi se nanašala na slovarje in besedišče, je bilo malo, kljub temu pa anketa prinaša nekaj rezultatov, relevantnih za obravnavano temo.

Učiteljem se učenje besedišča v procesu izobraževanja zdi zelo pomembno, zato bi v idealnih razmerah tej dejavnosti pri pouku posvetili veliko časa. Manj bi ga namenili spoznavanju jezikovnih priročnikov in njihovi rabi, čeprav se jim tudi to zdi razmeroma pomembno. Podobno učenci trdijo, da je velik besedni zaklad najpomembnejši element dobro razvite sporazumevalne zmožnosti,[8] poznavanje slovarjev se jim v primerjavi z ostalimi znanji in spretnostmi zdi manj pomembno, tudi od poznavanja interneta, ki so ga ocenili razmeroma visoko. V skladu s tem so tudi odgovori, ki so se nanašali na rabo različnih jezikovnih virov in informacijsko-komunikacijske tehnologije (IKT): učenci za reševanje jezikovnih težav redno uporabljajo elektronske vire, predvsem spletne iskalnike, bistveno več kot slovarje, kar še posebej velja za tiste v knjižni obliki, kar je v skladu z rezultati, predstavljenimi pod točko 2.1. Nasprotno pa učitelji (predvsem starejši) spletnih slovarjev in IKT pri pouku ne uporabljajo pogosto, čeprav so načelno temu naklonjeni.

2.4

V okviru projekta SSJ je bila izvedena tudi empirična raziskava, ki je preverjala razumljivost slovarskih informacij o slovničnih (oblikoskladenjskih) lastnostih besed (Rozman et al. 2010), v kateri je sodelovalo 389 učencev 8. in 9. razreda OŠ ter 2. in 3. letnika SŠ. Izkazalo se je, da so bila na novo napisana gesla, v

6 Učitelji uporabijo slovar kot vir informacij o pomenu besede ali pa, sicer redkeje, slovar učenci uporabijo sami.

7 http://www.slovenscina.eu/ (dostop 30. 6. 2015).

8 Vprašanje se je glasilo: Katera od spodaj napisanih znanj in spretnosti so po tvojem mnenju še posebej pomembna za to, da uspešno govoriš, pišeš in bereš v slovenskem knjižnem jeziku? Na voljo so imeli 8 odgovorov, vsakega so ocenili z oceno od 1 do 6.

katerih so bili podatki o slovničnih lastnostih besed kar se da eksplicitni, razumljivejša[9] od gesel iz SSKJ, kjer so ti podatki podani kot (kvalifikatorske) krajšave, v glavi in zaglavju pa zapisani zelo zgoščeno. Kot najkoristnejši za razumevanje so se izkazali tisti deli geselskih člankov, v katerih so bili podatki o slovnici bolj eksplicitni in najbolj neposredno povezani z vprašanjem, zastavljenim v testni nalogi, ne glede na to, na katerem mestu so bili ti podatki zapisani: razumljive in za reševanje testa koristne so bile tako razvezane krajšave kvalifikatorskih pojasnil, zapisane za iztočnico, kot grafično izpostavljena eksplicitna pojasnila in zgledi rabe, zapisani za definicijo.

2.5

V letih 2007–2009 je bila v okviru priprave doktorske disertacije (Rozman 2010) opravljena obsežna analiza učnih načrtov in učbenikov za 2. in 3. triletje OŠ in srednje šole, ki je preverjala vključenost slovarskih vsebin v jezikovni pouk slovenščine, ter bila izvedena empirična raziskava o razumljivosti slovarskih definicij, v kateri je sodelovalo 607 učencev iz treh starostnih skupin: 5./6. razred OŠ, 8./9. razred OŠ in 2./3. letnik SŠ.

Zmožnost uporabe slovarjev sodi med pričakovane dosežke ob koncu osnovnošolskega izobraževanja,[10] slovar je v učne načrte vključen od 7. razreda OŠ naprej, v šolskih gradivih pa se vaje za delo s slovarjem (predvsem SSKJ) pojavljajo že dosti prej. Analiza je izpostavila več težav pri obravnavi in vključevanju slovarjev v didaktični proces, ki med drugim izhajajo iz neupoštevanja dejstva, da je SSKJ v nekaterih segmentih močno zastarel, še bolj pa, da je zaradi količine, strukturiranosti in neeksplicitnosti podatkov o besedišču marsikdaj prezahteven za takšno uporabo, kot jo predvidevajo naloge.

Empirična raziskava se je osredotočila na primerjavo razumljivosti definicij iz SSKJ in novih definicij, ki so bile napisane posebej za namene testiranja, in sicer tako, da bi bile učencem čim bolj razumljive (upoštevale so načela eksplicitnosti, neposrednosti, izogibanja abstraktnim in strokovnim besedam, nezapletene stavčne strukture, manjše pomenske razdrobljenosti itd.). Predpostavka, ki se je s testiranjem potrdila, je bila, da so definicije iz SSKJ zaradi abstraktnosti in prezahtevnega definicijskega jezika težje razumljive, predvsem mlajšim učencem. Raziskava je tudi izpostavila nekaj lastnosti, ki vplivajo na razumljivost definicij in se nanašajo na (ne)posrednost, strukturo, dolžino in tip definicij ter dolžino in strukturo geselskih člankov.

9 Še posebej osnovnošolcem.

10 »Poimenovalno zmožnost [učenec/učenka] pokaže tako, da: /…/ zna uporabljati slovarske priročnike v knjižni in elektronski obliki (7., 8., 9. razred)« (UN za slovenščino v OŠ: 89).

2.6

Od leta 2010 obstaja prosto dostopen korpus šolskih pisnih besedil Šolar,[11] ki vsebuje avtentična besedila učencev, vključenih v slovenski šolski sistem, in je zato odličen vir informacij o njihovi pisni jezikovni zmožnosti. Izčrpna analiza Šolarja je bila narejena v okviru gradnje Pedagoškega slovničnega portala[12] (Kosem et al. 2012), vendar pa je za pripravo slovarskih vsebin ta analiza le deloma relevantna, saj so bile napake[13] kategorizirane glede na jezikovne zadrege (npr. napake zapisa, skladnje), zato informacij o problematičnosti posameznih besed ne daje. Leta 2015 je bila zato opravljena analiza, ki je bila usmerjena prav v pridobivanje podatkov za pripravo šolskega slovarja in didaktičnih gradiv za obravnavo besedišča (Arhar Holdt in Rozman 2015). Zajela je sicer le manjši del korpusa, potrdilo pa se je, da bodo na ta način pridobljene informacije dobro vodilo glede slovarske obravnave polnopomenskih, pa tudi slovničnih besed. Predvsem se je pokazala potreba po povezovanju tradicionalno slovarske in slovnične problematike, izpostavljanju kolokacijskih, stilističnih in sintagmatskih značilnosti besed, sopostavljanju oblikovno podobnih, a pomensko različnih besed ter prikazovanju skupnih značilnosti in posebnosti besed, ki so pomensko blizu ali delno prekrivne, raba pa je kolokacijsko, stilistično ali kako drugače omejena.

2.7

Vse zgoraj opisane raziskave so se osredotočale na učence in učitelje slovenščine v slovenskih osnovnih in srednjih šolah, slovarskih raziskav na področju slovenščine kot drugega in tujega jezika pa skoraj nimamo. Izjema je raziskava Rozman (2003), v kateri je bila opravljena analiza angleških enojezičnih slovarjev za tujce ter opravljena kratka anketa med 64 udeleženci in 18 učitelji tečajev slovenščine kot drugega in tujega jezika, ki so potekali poleti 2003 v organizaciji Centra za slovenščino kot drugi/tuji jezik (CSDTJ).[14] Na podlagi teh rezultatov je bila utemeljena potreba po pripravi enojezičnega slovarja slovenščine za tujce, podani pa so bili tudi nekateri predlogi za konkretne slovarske rešitve. Podobno, a krajšo anketo smo v okviru CSDTJ izvedli tudi maja 2015; na anketo z desetimi vprašanji je odgovorilo 15 učiteljev, ki poučujejo slovenščino kot drugi in tuji jezik.

Glavne ugotovitve obeh anket lahko strnemo takole:

- velika večina učečih se pri učenju slovenščine uporablja slovarje,
- pretežno uporabljajo dvojezične slovarje, ki so v paru z njihovim prvim jezikom,

11 http://www.slovenscina.eu/korpusi/solar (dostop 30. 6. 2015).
12 http://www.slovenscina.eu/portali/pedagoski-slovnicni-portal (dostop 30. 6. 2015).
13 Za napako se je v Šolarju štelo le tisto, kar so popravili učitelji.
14 http://www.centerslo.net/ (dostop 30. 6. 2015).

- od enojezičnih slovarjev uporabljajo le SSKJ,
- SSKJ uporabljajo učeči se z boljšim znanjem slovenščine in na višjih stopnjah učenja, govorci drugih slovanskih jezikov in jezikoslovno izobraženi,
- slovarje uporabljajo pri različnih dejavnostih, najpogosteje pri pisanju, prevajanju in branju,
- večina učečih se bi uporabljala enojezični slovar slovenščine za tujce;
- precej učiteljev pri pouku uporablja enojezične slovarje slovenščine, in sicer za različne dejavnosti, predvsem za prevajanje in delo z besediščem (iskanje pomenov, primerov rabe, frazemov, sopomenk, besednih družin ipd.),
- večina učiteljev meni, da bi potrebovali slovar za tujce,
- enojezični slovar bi tako lahko začeli uporabljati prej (že na nižjih stopnjah jezikovnega znanja, po mnenju učiteljev na stopnji A2–B1, SSKJ pa na stopnji B2 Skupnega evropskega jezikovnega okvira),[15]
- tak slovar bi bil bolj primeren za različne dejavnosti pri pouku, tudi za tvorjenje in razumevanje besedil,
- veliko učiteljev meni, da bi tak slovar predvsem moral vsebovati preproste razlage in čim več zgledov rabe ter biti v e-obliki,
- po mnenju učiteljev učeči se poleg pomena besed najpogosteje potrebujejo informacije o rabi in slovničnih lastnostih besed,
- pri svojih študentih pa pri uporabi besedišča najpogosteje opažajo napake v skladnji in pomensko ter kontekstualno neustrezno rabo.

3 UPORABNIŠKE RAZISKAVE V TUJINI

Pregled tujih raziskav o rabi slovarjev kaže ravno obratno stanje, kot je v Sloveniji, saj tam prevladujejo raziskave med tujimi govorci, raziskave med govorci prvega jezika pa so zelo redke. Poleg tega so skoraj vse raziskave osredotočene na študente, največkrat na študente tujih jezikov ali prevajalce, pa tudi jezikoslovce in učitelje jezikov. V nadaljevanju predstavljamo nekaj ključnih ugotovitev omenjenih raziskav, pri čemer so mnoge relevantne za snovalce slovarjev tako za domače kot tudi tuje govorce.

S tehnološkim napredkom so se mnoge raziskave osredotočile na preučevanje razlik med uporabo različnih slovarskih medijev. Že v 90-ih letih 20. stoletja je

15 Skupni evropski jezikovni okvir (http://www.europass.si/files/userfiles/europass/SEJO%20komplet%20za%20splet.pdf, dostop 30. 6. 2015) opisuje jezikovno zmožnost na šestih ravneh, pri čemer A2 pomeni drugo, B1 pa tretjo raven.

Leffa (1993) primerjal rabo elektronskih in tiskanih slovarjev pri osnovnošolcih in ugotovil, da so učenci prevajali besedila bolje in hitreje z elektronsko verzijo slovarja kot tiskano, poleg tega pa je 80 % učencev imelo elektronsko verzijo raje od tiskane. Podobno preferenco so pokazali tudi študenti španščine kot tujega jezika v raziskavi Aust et al. (1993), pri čemer je bila ena od izpostavljenih prednosti elektronskih slovarjev tudi ta, da so študenti lahko v elektronskem slovarju v istem času poiskali več besed kot v tiskanem. Do podobnih zaključkov so kasneje prišli tudi Nesi (2000), Corris et al. (2000), Tono (2000), Laufer (2000), Winkler (2001), Laufer in Levitzky-Aviad (2006), Petrylaite et al. (2008) in Dziemianko (2010). Nekatere od omenjenih raziskav so podale tudi druge relevantne ugotovitve. Tako je npr. Laufer (2000) v raziskavi s 55 študenti angleščine kot tujega jezika ugotovila, da so bili rezultati vaje iz razumevanja neznanih besed občutno boljši, ko je bila študentom ponujena kombinacija prevoda, razlage in primerov. Winkler (2001) je v svoji raziskavi ugotovila, da so spretnosti, ki jih mora uporabnik imeti pri uporabi elektronske in tiskane verzije slovarja, včasih zelo različne, kar velja tudi za težave, s katerimi se sreča pri uporabi različnih oblik slovarjev. Z vidika slovarskih medijev so relevantne tudi ugotovitve Chenove raziskave (2010) o rabi tiskanih in elektronskih žepnih[16] slovarjev. 85 udeležencev raziskave, Kitajcev, ki so se učili angleščine, je veliko pogosteje uporabljajo elektronske žepne slovarje, je bila pa opažena razlika v rabi zaradi količine podatkov na strani tiskanega slovarja oz. zaslonu elektronskega žepnega slovarja – uporabniki so raje uporabljali elektronske slovarje pri branju, tiskane pa pri prevajanju in pisanju.

Mnoge raziskave poročajo o tipih informacij, ki jih slovarski uporabniki največkrat iščejo. Po pričakovanju sta na prvem mestu razlaga in zapis (Béjoint 1981; Jackson 1988; Battenburg 1989; Harvey in Yuill 1997; Hartmann 1999; Kosem 2010; Verlinde in Binon 2010; Lorentzen in Theilgaard 2012), pogosto pa tudi sinonimi. Tuji govorci pogosto poiščejo tudi slovnične informacije, informacije o kolokacijah, zglede in idiome oz. frazeologijo (Béjoint 1981; Harvey in Yuill 1997). Ostale informacije, kot sta npr. etimologija in izgovorjava, uporabniki iščejo zelo redko (Hartmann 1999; Kosem 2010). Posebej velja izpostaviti Kosmovo (2010) obsežno raziskavo med 444 domačimi govorci in 169 tujimi govorci angleščine na Univerzi Aston, katere rezultati so prikazani v Tabeli 3 in kažejo, da tuji govorci skoraj vse tipe informacij, z izjemo zapisa, iščejo pogosteje kot domači govorci, pri čemer velja upoštevati dejstvo, da je nekaterim tipom informacij (npr. zgledom, kolokacijam) v slovarjih za tujce namenjeno več poudarka.

16 Žepni slovarji so slovarji manjšega obsega (manjše število gesel in kratke, poenostavljene informacije o iztočnicah).

Tabela 3: Uporaba mikrostrukturnih delov slovarskega gesla (domači in tuji govorci) (vrednosti: 1 – skoraj nikoli, 2 – redko, 3 – pogosto, 4 – skoraj vedno); prevedeno iz Kosem (2010: 162).

	domači govorci (povprečje)	tuji govorci (povprečje)
razlage	3,44	3,56
zapis	2,82	2,73
sinonimi	2,63	2,91
zgledi	2,45	2,92
raba in slovnica	1,72	2,16
frazeologija	1,66	2,27
kolokacije	1,49	2,15
izgovorjava	1,60	2,10

Obstaja tudi kar nekaj raziskav o tem, katere besede uporabniki iščejo. V Béjointovi (1981) raziskavi 66 % študentov (tujih govorcev) sploh nikoli v slovarju ni iskalo pogostih besed, podobno pa so kasneje potrdili Hatherall (1984), Bogaards (1998) ter Nesi in Haill (2002). Teh ugotovitev nista potrdila Verlinde in Binon (2010), ki sta pri analizi 55.752 dostopov do spletnega portala Base lexical du français (BLF) ugotovila, da uporabniki pogosto iščejo pogoste besede. Podobno so pri pregledu približno pol milijona dostopov svahilsko-angleškega slovarja ugotovili tudi de Schryver et al. (2006), ki pa hkrati zaključujejo, da obstaja določena korelacija med pogostostjo besed v korpusu in pogostostjo njihovih poizvedb zgolj za prvih nekaj tisoč besed in je ravno zato nemogoče predvideti, katere besede bodo zanimale slovarske uporabnike. Tudi Trap-Jensen et al. (2014) so pri analizi dnevnikov iskanj uporabnikov spletne verzije danskega slovarja (Der Danske Ordbog)[17] ugotovili, da so funkcijske in korpusno najpogostejše besede med najpogosteje iskanimi besedami (60 % od 1.000 najbolj iskanih besed najdemo med 10.000 najpogostejšimi besedami v korpusu).

Raziskave so razkrile nekaj strategij uporabe slovarjev, ki pogosto odločajo o tem, kako uspešni so uporabniki pri iskanju in interpretaciji informaciji v slovarjih. Med pogosto omenjenimi je ogled samo prve razlage oz. razlage pri prvem pomenu, o kateri poročajo Mitchell (1983), Tono (1984), Neubach in Cohen (1988), McCreary (2002), Nesi in Haill (2002) in Kosem (2010). Podobno za elektronske žepne slovarje ugotavlja tudi Boonmoh (2012), pri čemer so študenti v raziskavi pogledali samo del gesla, ki je bil prvotno prikazan na zaslonu. S tega vidika je zelo pomembna odločitev o vrstnem redu pomenov v slovarskih geslih, ki mora

17 http://ordnet.dk/ddo (dostop 30. 6. 2015).

biti prilagojen ciljnemu uporabniku, ter o izbiri različnih strategij za podajanje informacij, kot je npr. pomenski meni na začetku gesla za hiter pregled pomenske členitve in hitrejšo navigacijo po geselskem članku. Druga strategija, zaznana tako pri domačih kot tujih govorcih, je strategija »kidrule«, pri kateri uporabnik znano besedo iz razlage uporabi kot enakovrednico neznane besede. Strategijo sta prvič zaznala Miller in Gildea (1987) pri raziskavi, ki je vključevala 10 in 11 let stare otroke, kasneje pa so jo pri študentih in odraslih potrdili še Harvey in Yuill (1997), McCreary in Dolezal (1999), Nesi (2000), McCreary (2002) ter Nesi in Haill (2002). V tretjo skupino lahko uvrstimo primere, ko uporabniki, tudi zaradi neustrezno uporabljenih strategij iskanja, naletijo na težave pri rabi slovarjev. Tako recimo Selva in Verlinde (2002) poročata o težavah uporabnikov pri iskanju pravih informacij v večpomenskih geslih in v dolgih razlagah, podobne težave uporabnikov je pri obsežnejših geslih zaznal tudi Tono (2011).

Na splošno so pri raziskavah največkrat uporabljeni slovarji za tuje govorce angleščine,[18] zlasti zato, ker so ravno ti slovarji gonilna sila leksikografskih inovacij in posledično zelo zanimivi za odkrivanje novih trendov v rabi slovarjev. Med inovacijami, ki so jih v leksikografijo uvedli ti slovarji, so razlagalno besedišče, tj. omejen nabor besed, ki se lahko pojavijo v slovarskih razlagah (prvič uporabljeno v Longman Dictionary of Contemporary English), celostavčne razlage (prvič v slovarju COBUILD), pomenske indikatorje (Longman Dictionary of Contemporary English), pomenske menije (Macmillan English Dictionary for Advanced Learners) ter vpeljavo podatkov iz korpusov usvajanja jezika.[19] Glavni namen naštetih inovacij je predvsem olajšati uporabniku iskanje relevantne informacije ter njeno uporabo za namene tvorjenja besedil. Ravno zato ni presenetljivo, da celostavčne razlage, indikatorje in pomenske menije danes najdemo tudi v slovarjih za domače govorce.

V zvezi s slovarji v tujini velja omeniti še en pomemben trend. Slovarske vsebine se selijo na splet in dejansko težko najdemo sodobni slovar, ki ne bi imel tiskane in spletne verzije. Vse pogosteje se celo dogaja, da slovarji sploh nimajo več tiskanih verzij (npr. znani angleški založnik Macmillan je leta 2012 ukinil izdelavo tiskanih slovarjev in se posvetil samo spletnim slovarjem (gl. Rundell 2014). Vendar pa so spletne strani slovarjev postale mnogo več, dejansko gre za portale z več referenčnimi viri (poleg slovarjev najdemo tudi tezavre itd.) in z različnimi informacijami o jeziku, npr. z blogi o določenih posebnostih rabe besed, opozorili na najpogostejše jezikovne napake, multimedijskimi vsebinami ipd. Na ta način slovar postaja zgolj del jezikovno-didaktičnega servisa. Za šolsko rabo je zanimiv portal Wordsmyth,[20] katerega osrednji del so otroški, ilustrirani in šolski slovarji

18 Predvsem tisti, ki so namenjeni učečim se z boljšim jezikovnim znanjem (od ravni B2 naprej).
19 Gre za korpuse besedil, ki so jih napisali tuji govorci.
20 http://www.wordsmyth.net (dostop 30. 6. 2015).

za domače govorce angleščine, poleg tega pa uporabniki na strani lahko uporabijo orodja za reševanje anagramov in ugank ter izdelavo glosarjev in kvizov. Takšna orodja so dobrodošla tako z vidika učencev kot učiteljev, saj olajšajo vključevanje slovarskih vsebin v pouk.

4 ZAKLJUČKI IN VIZIJA

Kaj torej vemo o potrebah, zmožnostih in navadah uporabnikov in kako naj to upoštevamo pri pripravi slovarskih virov?

4.1

Raba in poznavanje slovarjev sta del predpisanih kurikularnih vsebin, slovarske vsebine se pojavljajo v šolskih gradivih, učitelji pa slovarjev (predvsem SSKJ) ne uporabljajo le kot priročnikov za pripravo učnih gradiv in vrednotenje učenčevih izdelkov, ampak tudi med poukom v razredu. Sodeč po rezultatih anket, slovarje med poukom uporabljajo pri obravnavi različnih jezikovnih vsebin in pri različnih, predvsem produktivnih dejavnostih. Slovar v šolski praksi torej ni le priročnik, ki daje informativno-normativne podatke o jeziku, ampak je tudi pomembno didaktično sredstvo. Domnevamo lahko, da bi bil slovar, ki bi bil narejen z mislijo na upoštevanje potreb in zmožnosti učencev ter zasnovan kot priročnik za delo v razredu, za pouk še bolj privlačen, predvsem pa bi sodoben deskriptivni slovar s poudarjeno didaktično vlogo verjetno pomembneje pozitivno vplival na razvoj prožne sporazumevalne zmožnosti. Zdi se, da je za to potrebno preseči tradicionalne okvire slovarske forme in razmišljati o izdelavi spletnega portala, kjer bo na enem mestu zbranih več vrst leksikalnih informacij. Kot nakazuje analiza Arhar Holdt in Rozman (2015),[21] bo potrebno, tako s pomočjo korpusnih analiz kot na podlagi analiz pouka slovenščine, detektirati leksikalne težave učencev in nanje v slovarju ne samo opozarjati, ampak jih tudi reševati. To pomeni, da klasične slovarske informacije dopolnjujemo z eksplicitnimi (oz. bolj »šolskimi«) razlagami posameznih vsebin, jih v posebnih razdelkih povezujemo s slovničnimi, pravopisnimi idr. jezikovnimi vsebinami, opisujemo stilistične, kolokacijske, sintagmatske razlike sopomenk ali delno pomensko prekrivnih besed, sopostavljamo in opozarjamo na različne pomene in rabe oblikovno podobnih besed, na posebne rabe, ki odstopajo od usvojenih vzorcev, dajemo normativne nasvete, jih povezujemo s kvizi, vajami, dopolnjujemo s slikami, seznami pogostih napak, besede vključujemo v pomenska polja in besedotvorne družine, združujemo informacije o besedah na podlagi različnih asociativnih povezav itd. S takimi povezavami

21 In podobne so prakse tudi v tujini, prim. npr. Vocabulary.com, Merriam-Webster.com.

bodo učenci nove informacije o besedišču usvajali uspešneje, saj jih bodo lažje povezali z že usvojenim znanjem in jih vključili v mentalni leksikon, kjer so besede med seboj po obliki in vsebini povezane v leksikalne mreže (prim. npr. Rozman 2010: 32), poleg tega pa bomo z združevanjem različnih informacij o besedah in vaj spodbujali tudi razvoj strategij za uspešno učenje besed (prim. Paynter et al. 2005: 30–68). Eksplicitna pojasnila o zadregah glede norme in rabe, ki presegajo v šolah še vedno pogosto prisotno dekontekstualizirano črno-belo slikanje jezikovnih pojavov na prav-narobe (Stabej et al. 2008), pa so izjemno pomembna za razvoj boljšega razumevanja kompleksnega delovanja jezika (in s tem prožne sporazumevalne zmožnosti posameznikov).

4.2

S takšno informativno-didaktično naravnanostjo, kot jo opisujemo v razdelku 4.1, bi slovar verjetno naredili učencem privlačnejši tudi za samostojno rabo. Slovarjev, tudi spletnih, učenci – kot so pokazale raziskave – trenutno skoraj ne uporabljajo, čeprav je iskanje informacij po spletu (preko spletnih iskalnikov) običajen način pridobivanja informacij o jeziku. Natančnejših vzrokov za to ne poznamo, domnevamo pa, da vzroki tičijo tudi v neprijaznosti spletnih slovarskih vmesnikov in uporabniku neprijazni obliki slovarskih sestavkov z zelo zgoščenimi, nepregledno strukturiranimi ter težko razumljivimi slovarskimi podatki.

Empirični raziskavi, predstavljeni pod razdelkoma 2.4 in 2.5, ki sta se sicer osredotočali samo na določene segmente slovarja, sta namreč pokazali, da je SSKJ (kot najbolj uporabljan slovar pri pouku slovenščine) na več mestih težko razumljiv, predvsem mlajšim učencem.

Izkazalo se je, da so zakodirane informacije o slovničnih lastnostih besed slabo razumljive in na podlagi rezultatov lahko zaključimo, da je bolj eksplicitno načeloma boljše, pa naj bo zapisano kot kvalifikatorsko pojasnilo, izpostavljeno v zgledu rabe ali v posebnem grafično izpostavljenem razdelku o rabi besed. Za določitev, katere informacije o slovničnih lastnostih besed je potrebno v slovarju bolj eksplicirati in kako, bi sicer potrebovali še kakšno obširnejšo raziskavo, ki bi jasneje izpostavila problematična mesta in na podlagi katere bi lahko določili najustreznejše načine podajanja informacij o slovnici (in tudi drugih lastnostih besed, npr. o stopnji formalnosti, figurativni rabi, kolokativnosti).

Pri pisanju definicij pa velja, da je upoštevanje spoznavno-jezikovnega razvoja učencev izjemno pomembno, saj so se v raziskavi na novo napisane definicije, ki so upoštevale spoznanja o razvoju abstraktnega mišljenja ter posameznih delov

sporazumevalne zmožnosti,[22] izkazale za bistveno razumljivejše od definicij iz SSKJ. Upoštevati je namreč treba, da se v času šolanja število in tipologija usvojenih besed ter razumevanje abstraktnih pomenov, odnosov med različnimi besedami oz. pomeni, daljših in zapletenih stavčnih struktur, oblikoslovnih in besedotvornih pravil itd. z leti povečuje, deloma zaradi spoznavnega in socialnega razvoja, deloma zaradi šolanja oz. vpliva jezikovnega pouka (prim. Rozman 2010). Ker pa se je pisanje novih definicij v omenjeni raziskavi osredotočalo zgolj na čim večjo razumljivost, ne daje odgovora na (za slovaropisje še kako pomembno) vprašanje, ali so definicije, primerne za mlajše, sprejemljive in ustrezne tudi za starejše učence.

Raziskava je tudi izpostavila nekaj značilnosti, ki vplivajo na razumljivost definicij (npr. neposredne definicije so boljše od posrednih; definiranje s pogosto rabljenimi besedami je načeloma dobro, slabo pa v primeru, ko so besede zaradi večpomenskosti lahko manj natančne in slabše predstavne; definicije naj ne vsebujejo manj rabljenih strokovnih besed, naj ne bodo preabstraktne in pomensko preveč zgoščene), nakazane pa so bile tudi možnosti členjenja pomenov pri večpomenskih besedah (npr. pomen besede v danem kontekstu je zelo težko razumeti, če so različni pomeni pojasnjeni z več med sabo močno (pomensko in besedilno) povezanimi, zgoščenimi in abstraktnejšimi definicijami), kar velja pri snovanju slovarja za šolsko publiko upoštevati.

4.3

Že opravljene raziskave dajejo določene smernice za oblikovanje slovarja, ki bo ustrezen za učence in primeren tudi za uporabo med poukom. Kljub temu pa ostaja še kar nekaj odprtih vprašanj. Eno izmed njih se nanaša na obravnavo slovničnih besed, s katerimi se raziskave niso dosti ukvarjale, analiza, opisana v Arhar Holdt in Rozman (2015), pa nakazuje potrebo po razširitvi klasične slovarske obravnave s slovnično.

Prav tako nam manjkajo empirični podatki, na katere bi se lahko naslonili pri izbiri geslovnika. Iz teorij o usvajanju besedišča vemo, da se v času šolanja začne besednjak širiti predvsem z večzložnimi, abstraktnimi in strokovnimi besedami, kasneje tudi z manj pogostimi in tistimi, vezanimi na ožja področja. Vendar omejitev le na ta segment besedišča najbrž ni smiselna, saj analize korpusa Šolar kažejo (Kosem et al. 2012, Arhar Holdt in Rozman 2015), da je, vsaj na ravni produkcije, veliko težav povezanih prav s splošnejšim besediščem.[23] Zato velja nadaljevati s korpusnimi in drugimi raziskavami leksikalne problematike, poleg tega pa bi

22 Gl. razdelek 2.5.

23 Podobno tudi mednarodne raziskave ne dajejo enoznačnega odgovora o tem, katere besede s stališča pogostosti rabe (pogosto ali redkeje rabljene) uporabniki v slovarjih iščejo pogosteje (gl. razdelek 3).

kazalo narediti tudi sezname besedišča, rabljenega v šolskih gradivih in literaturi, morda celo v šolskem diskurzu, saj se sporazumevalna zmožnost razvija pri vseh šolskih predmetih. Brez analiz se zdi, da je vsaj s stališča rabe besed smiselno razmišljati o vključevanju tiste leksike, ki izkazuje različne pomenske prenose, nenavadne rabe oz. odstopanja od vzorca, kjer prihaja do dvojnic, pomenskih in oblikovnih podobnostih ipd., torej besedišče, pri katerem lahko – tudi na podlagi drugih analiz jezikovnih težav govorcev[24] – predvidimo težave.

4.4

V razdelkih 4.1–4.3 smo se osredotočali na šolske uporabnike s slovenščino kot prvim jezikom.[25] Kaj pa lahko rečemo o tujih uporabnikih in slovarju slovenščine za tujce? Zagotovo je potrebno izpostaviti, da tuji govorci, ki se uvrščajo med potencialne uporabnike slovarja, niso homogena skupina – razlikujejo se po svojem prvem jeziku in stopnji znanja slovenščine, nekateri se slovenščino učijo na tečajih, drugi v okviru študija, nekateri v Sloveniji, drugi v tujini, nekateri se v organizirane oblike izobraževanja niti ne vključujejo. Učeči se slovenščine kot tujega jezika se razlikujejo tudi po svojih potrebah in motivaciji za učenje; z njimi so povezani tudi različni interesi, ki jih imajo pri učenju, ter cilji, ki jih želijo doseči. Vse te razlike vplivajo na t. i. uporabniške potrebe, kljub temu pa med njimi obstaja dovolj skupnih točk, ki lahko predstavljajo osnovo pri snovanju slovarjev za tuje govorce. To nenazadnje dokazuje tudi obsežen in uspešen trg angleških slovarjev za tujce, iz katerih, kot je bilo rečeno že v razdelku 3, inovacije črpajo celo snovalci slovarjev za domače govorce. Slovarji za tuje govorce angleščine (kot tudi rezultati mednarodnih raziskav, opravljenih za to ciljno publiko) so vsekakor dober vir informacij tudi za pripravo slovarjev slovenščine, primernih tako za tuje govorce kot domače šolske uporabnike. Kljub temu pa obstaja vrsta vprašanj, specifičnih za slovenščino in slovensko situacijo, in ta so v marsičem podobna tistim, ki se zastavljajo za šolske uporabnike in o katerih smo pisali v prejšnjih točkah tega razdelka. Vendar pa imamo odgovorov za to ciljno publiko še bistveno manj, ker imamo opravljenih precej manj raziskav in ker še nimamo izdelanega dovolj velikega korpusa usvajanja slovenščine,[26] ki bi omogočal empirične raziskave pisne (pa tudi govorne) produkcije tujih govorcev slovenščine.

Prepotrebnih nalog za slovensko leksikografijo je torej dovolj – pri tem se lahko opre na kar dobro razvito področje poučevanja slovenščine kot drugega in tujega jezika.

24 Prim. npr. Bizjak Končar et al. (2011).

25 Čeprav je seveda to poenostavljen prikaz – predstavljene raziskave, na podlagi katerih smo izpeljali zaključke, so se nanašale na učence (in učitelje) v slovenskih šolah, kjer je seveda tudi precej takih, ki jim slovenščina ni prvi jezik.

26 Zaenkrat imamo izdelan le manjši korpus brez označenih napak, ki vsebuje 32.117 besed oz. 306 besedil (Rozman et al. 2010) ter poskusni korpus piKUST v skupnem obsegu 34.873 besed oz. 128 pisnih besedil ter z označenimi napakami, ki jih je skupaj 5.085 (Stritar 2012).

Tu so se uveljavile različne metode poučevanja, pomnjenja in učenja slovenske leksike, ki so do neke mere tudi dokumentirane v učbenikih in različnih dodatnih didaktičnih gradivih.[27] Vedenje o tem, katero besedišče in na kakšen način se posreduje tujim govorcem slovenščine, bi zagotovo lahko bilo v pomoč pri snovanju slovarja kot didaktičnega pripomočka v procesu učenja slovenščine kot drugega in tujega jezika, seveda pa tudi pri pripravi geslovnika. Geslovnik bi se dalo pripraviti s pomočjo seznamov besed, ki jih vsebujejo učbeniki in druga gradiva, npr. dokument Sporazumevalni prag za slovenščino (Ferbežar et al. 2004), ki opisuje znanje slovenščine na ravni B1 Skupnega evropskega jezikovnega okvira. Sporazumevalni prag bi bil pri snovanju slovarja uporaben tudi zato, ker je v njem besedišče zbrano po tematskih sklopih (t. i. posebni pojmi) in kategorizirano v smislu časovnih, prostorskih, količinskih idr. pojmov (t. i. splošni pojmi), kar je za tujega uporabnika lahko uporabna dodatna informacija. Pri snovanju geslovnika pa bi kazalo razmisliti tudi o uporabnosti besedišča, ki se uporablja v izpitnih polah, namenjenih preverjanju znanja slovenščine kot drugega in tujega jezika na različnih ravneh.

5 ZA KONEC

Po pregledu raziskav o rabi slovarjev ter razmisleku o slovarskih potrebah učencev v slovenskih šolah in tujih govorcev se zastavlja vprašanje, ali je možno izdelati en sam slovar, ki bi bil namenjen tako tujim govorcem kot šolskim uporabnikom. Vprašanje se zdi zanimivo z vidika uporabe v okviru izobraževanja oz. pri učenju jezika. Procesa usvajanja besedišča prvega in tujega jezika sta sicer v marsičem drugačna, hkrati pa imata precej skupnih točk (Singleton 1999: 79–82; prim. tudi Jesenovec 2004). Katere skupne točke lahko ujamemo v slovar, potrebuje ne samo temeljit razmislek, ampak tudi kako obsežno raziskavo, zagotovo pa je pri načrtovanju slovarja (oz. slovarjev) za obe publiki nujno razmišljati o poudarjeni didaktični vlogi – in tukaj so slovarske rešitve močno prekrivne; tisto, kar smo za šolski slovar ugotavljali v razdelku 4.1, v veliki meri velja tudi za slovar za tujce.

Po drugi strani se je potrebno vprašati, ali bi bil sodoben splošni slovar lahko primeren tudi za učence in tujce. To je sicer nekoliko v nasprotju s temeljno mislijo našega prispevka, vendar pa nam povsem negativen odgovor preprečuje nekaj pomislekov. Prvi je ta, da so se vse raziskave, ki so preverjale rabo splošnega slovarja, nanašale na SSKJ, ki je v marsikaterem pogledu že precej zastarel, nekatere analize pa kažejo, da je težje razumljiv celo odraslim domačim govorcem slovenščine[28] (Kosem 2006: 26; gl. tudi Müller 1996 in 2009). Drugi pomislek so mnenja učiteljev, da je uporaba splošnega slovarja primerna za ti dve ciljni publiki; tudi če se s tem ne strinjamo, najbrž ne moremo povsem zanikati, da je uporaba takšnega

27 Npr. v različnih priročnikih k učbenikom, dostopnih na spletnih straneh CSDTJ.

28 Čeprav empiričnih dokazov za to nimamo, saj raziskav, ki bi preverjale razumljivost SSKJ pri odraslih, ni.

slovarja v določenih segmentih pouka lahko koristna. In nenazadnje, splošni slovar, ki bi upošteval tudi spoznanja o potrebah učencev in tujih govorcev ter bil narejen po sodobnih principih, upoštevajoč prevladujoče načine iskanja jezikovnih informacij ter pereče jezikovne zadrege govorcev slovenščine, bi bil morda veliko primernejši kot SSKJ. In oblika spletnega portala celo omogoča, da bi tak slovar vključeval tudi didaktične vsebine, ki so za druge uporabnike manj zanimive.

Na omenjeni vprašanji je težko odgovoriti z jasnim da ali ne, tudi zaradi dejstva, da trenutna slovarska oz. leksikografska situacija pri nas ne ponuja dovolj empiričnih dokazov, na katerih bi lahko utemeljili eno ali drugo odločitev. Ob delni prekrivnosti potreb obeh tipov uporabnikov (domačih in tujih govorcev) in možnostih kombiniranja leksikografskih rešitev, ki nam jih ponujajo sodobni mediji, je mogoče smiselno razmišljati o skupnem snovanju slovarske baze, ki bi vsebovala tako informacije, relevantne za vse tipe uporabnikov, kot tudi informacije, relevantne za posamezne tipe uporabnikov. Na ta način lahko kasneje izdelamo bodisi slovarje za posamezne tipe uporabnikov bodisi portal s podatki za vse tipe uporabnikov in različnimi vizualizacijami le-teh, lahko pa tudi oboje ali celo kaj tretjega, če se izkaže, da obstajajo boljše rešitve. Zato pa se je treba čim prej lotiti leksikografskega dela in raziskovanja, ki bo dalo čim več empiričnih informacij, na katerih bodo leksikografske rešitve osnovane.

Vloga jezikovnih vprašanj prevajalcev pri načrtovanju novega enojezičnega slovarja

Jaka Čibej, Vojko Gorjanc in Damjan Popič

Abstract
This paper focuses on dictionary use among translators. It begins with an overview of related work in this field and continues by presenting the results of a pilot study into translators' use of language resources when solving language problems. Through analysis of a number of discussions conducted in *Prevajalci, na pomoč!*, a dedicated, self-managed Facebook group for Slovenian translators aimed at solving translation problems, a taxonomy of typical language problem scenarios was developed. Based on the analysis of the problems encountered by translators, and their suggested solutions, we aim to establish the areas in which problems occur and the ways solutions are found. In the final segment, we discuss the suitability of the proposed method and the degree to which this approach yields results that are of use for the compilation of monolingual dictionaries.

Keywords: monolingual dictionary, language resources, dictionary users, translators

Ključne besede: enojezični slovar, jezikovni viri, slovarski uporabniki, prevajalci

1 UVOD

Slovarski uporabniki glede na svoje poklicno, jezikovno ali kulturno ozadje v slovarju iščejo in pričakujejo različne slovarske informacije, njihova pričakovanja pa se od skupine do skupine lahko precej razlikujejo. Pomembno je torej, da v raziskavo uporabniških pričakovanj in potreb (potencialnih) slovarskih uporabnikov zajamemo več raznorodnih skupin (od prostočasnih do poklicnih jezikoslovcev), saj na ta način dobimo karseda celovito in reprezentativno sliko pričakovanj, ki bi jih moral izpolniti nov enojezični slovar, zasnovan v skladu s sodobnimi leksikografskimi smernicami. Ena od omenjenih skupin poklicnih jezikoslovcev so tudi prevajalci, ki pri svojem delu uporabljajo različne jezikovne vire – med njimi tudi vrsto enojezičnih, npr. slovarje, korpuse, glosarje in slogovne priročnike (Hirci 2003).

Dosedanje raziskave na področju uporabe slovarjev pri prevajalcih so se veliko ukvarjale s pedagoško prakso, predvsem z uporabo slovarjev pri poučevanju prevajalcev ali z vključevanjem učenja uporabe slovarjev v pedagoški proces (Roberts 1992; Sanchez Ramos 2005; Hirci 2013), prav tako pa ni zanemarljiv delež raziskav o uporabi slovarjev pri profesionalnem prevajalskem delu (Nuccorini 1992; Roberts 1997; Atkins in Varantola 1998; Varantola 2002; Sanchez Ramos 2005). Če želimo, da bo novi enojezični slovar slovenščine karseda široko uporaben in uporabniku prijazen tudi za to specializirano skupino slovarskih uporabnikov, je nujno proučiti njihove potrebe in pričakovanja.

V pričujočem prispevku na kratko predstavljamo pregled dosedanjih uporabniških raziskav ter prve izsledke pilotne raziskave, v kateri smo se osredotočili na poklicne prevajalce in poskušali ugotavljati njihove potrebe z zbiranjem avtentičnih vprašanj v aktivni in angažirani skupnosti prevajalcev na Facebooku (skupina *Prevajalci, na pomoč!*). Na podlagi izsledkov bomo podali nekaj predlogov, ki bi jih bilo smiselno upoštevati ne le pri načrtovanju novega enojezičnega slovarja, temveč tudi pri morebitnih nadaljnjih načrtih za specializirane jezikovne vire, namenjene študentom prevajanja in poklicnim prevajalcem.

2 KRATEK PREGLED PODROČJA

Raziskave pri prevajalcih in študentih prevajanja se osredotočajo na uporabo slovarjev na eni strani in uporabnike na drugi, temu primerno pa je prilagojena tudi metodologija. V začetnih raziskavah uporabnikov je bil najpogosteje uporabljen metodološki pristop anketiranje, pri procesu prevajanja pa metoda TAP (angl. *think-aloud protocol*, protokol glasnega razmišljanja), pri kateri prevajalec med prevajanjem glasno ubeseduje svoje dileme in odločitve. Pri uporabnikih še

danes večinoma prevladuje anketiranje, za ugotavljanje uporabe slovarjev pa je metodološki spekter večji. Metoda TAP velja za zastarelo, saj lahko danes prevajalsko delo spremljamo tudi npr. s snemanjem računalniškega zaslona, z uporabo specializiranih programskih orodij (npr. Translog[1]), bolj kompleksno zastavljene raziskave pa danes uporabljajo tudi metodo sledenja pogledu (angl. *eye-tracking*) (Hirci 2009; Tono 2011). Raziskave uporabe jezikovnih virov vsaj pri prevajalcih večinoma niso več izolirane, temveč so del širših interdisciplinarno zasnovanih raziskav prevajalskega procesa (Paulsen Christensen 2011).

2.1 Raziskave v tujini

Raziskave uporabe slovarjev so se pri študentih zelo pogosto ukvarjale z njihovo uporabo pri učenju tujega jezika, še posebej pri učenju angleščine kot tujega jezika (Béjoint 1981; Mackintosh 1998; Humblé 2001). Nekakšno logično nadaljevanje tovrstnih raziskav so bile raziskave o uporabi slovarjev pri učenju prevajanja (Roberts 1992; Sanchez Ramos 2005; Hirci 2013), ki se večinoma ukvarjajo z različnimi vidiki uporabe slovarjev pri učenju jezika ali učenju prevajanja, predvsem skozi analizo uporabniške izkušnje ter s tem povezanimi načrti za bolj sistematično vključevanje slovarskih informacij v proces učenja.

Z vidika potreb med študenti prevajalstva, ki so jih zaznale tuje raziskave, je mogoče opaziti splošen trend: bolj kot je študent prevajalstva izkušen, pogosteje posega po enojezičnih slovarjih z obsežnejšimi podatki o pomenu in jezikovni rabi, na začetku pa se ponavadi zadovolji z informacijo iz dvojezičnega slovarja, v katerem išče v glavnem izolirane pomene določenega leksema (Sanchez Ramos 2005).

Raziskave uporabe slovarjev v prevajanju pa razkrivajo vrsto skupnih značilnosti, ki so aktualne predvsem za tiste, ki načrtujejo oz. pripravljajo jezikovne vire. V tem kontekstu lahko prepoznamo nekaj pogostih vzorcev, malodane aksiomov (Varantola 2002: 34):

- uporabniki slovarje uporabljajo za reševanje zadreg, ki so vezane na kontekst;
- uporabniki iščejo ustreznice v drugih jezikih, iščejo pa tudi potrditev, zato neradi vidijo ustreznice, ki jih ne poznajo;
- uporabniki potrebujejo tudi informacije o daljših segmentih besedila, ne zgolj o posameznih leksikalnih enotah;
- uporabniki v slovarskih virih iščejo tudi neslovarske informacije, saj te drugod niso zbrane in sistematično predstavljene.

1 http://www.translog.dk/ (dostop 1. 8. 2015).

Od naštetega lahko za pričujoči prispevek kot posebno pomemben vidik izpostavimo predvsem odvisnost od konteksta, saj se ta neposredno nanaša na uporabnost slovarske informacije. Ena temeljnih lastnosti uporabnosti slovarja pa je raven (samo)zaupanja, s katerim uporabnik neko slovarsko informacijo uporabi v konkretni situaciji (Varantola 2002: 34). To zelo poenostavljeno pomeni, da se prevajalci kot poklicni jezikovni uporabniki stežka zadovoljijo z okrnjenim naborom informacij v slovarju – da torej potrebujejo čim več informacij. Zavedati se moramo, kako pomembno je, da je slovarski opis zasnovan tako, da lahko na določenih mestih med ponujenimi informacijami prevajalci izbirajo sami, kar naj bi sicer bilo sploh vodilo dobrega slovarja (Atkins 1996: 532).

2.2 Raziskave v slovenskem prostoru

Čeprav raziskav o uporabi slovarjev v slovenskem prostoru ni veliko, je nekaj takih, ki se dotikajo prevajalcev v okviru študentske populacije oz. študijskega procesa. Gorjanc (2014) je npr. izvedel anketo, ki razkriva pričakovanja študentov prevajalstva glede idealnih enojezičnih slovarjev[2] in opozarja, da so študenti prevajalstva osveščeni glede sodobnih tehnoloških možnosti in da imajo v tem smislu do slovarja visoka pričakovanja, poleg tega pa so popolnoma domači v digitalnem okolju, kjer je slovarska informacija le ena od mnogih. Tipična lastnost dobrega slovarja je zato povezljivost slovarskih informacij z drugimi dostopnimi informacijami, s čimer slovar postaja le eden od jezikovnih virov, ki se uporablja v soodvisnosti od vsega drugega, kar je prosto dostopno in vedno na voljo. Pomembno je tudi, da slovarska informacija ni statična, temveč se stalno posodablja in nadgrajuje.

V marsičem podobne rezultate predstavlja tudi Hirci (2013) z raziskavo, ki je dragocena tudi zato, ker je bila izvedena v dveh časovnih obdobjih, in sicer v letih 2005 in 2012. Ta raziskava kaže na veliko spremembo pri uporabi slovarjev in drugih slovarskih virov pri študentih prevajanja v manj kot desetletnem obdobju: z uporabe bolj tradicionalnih slovarskih virov so se študentje popolnoma preusmerili na splet. Referenčni slovarski priročniki so bistveno manj uporabljani oziroma se praktično ne uporabljajo, če niso prosto dostopni v spletni različici. Mnogo bolj se uporabljajo zelo različni prosto dostopni viri, močno pa se je povečala tudi uporaba korpusov. Novost je uporaba blogov, na katerih študentje iščejo odgovor pri prevajalski skupnosti.

3 POTREBE POKLICNIH PREVAJALCEV: PILOTNA RAZISKAVA

Odločili smo se, da opravimo pilotno raziskavo o potrebah poklicnih prevajalcev, kot se ta odraža v diskurzu med samimi prevajalci. Slovenska prevajalska skupnost je

2 Anketa je bila izvedena med študenti prevajalstva in novinarstva, zato moramo biti glede posplošitev na prevajalce previdni.

namreč dobro samoorganizirana, tudi ko gre za uporabo sodobnih družbenih medijev. Veliko prevajalcev družbene medije uporablja vsakodnevno, večinoma za razreševanje konkretnih prevajalskih zadreg, zato smo predpostavljali, da bomo s sistematično analizo zastavljenih vprašanj znotraj prevajalske skupnosti lahko prišli tudi do podatkov, na kakšen način se pristopa k reševanju posameznega prevajalskega problema ter v kolikšni meri rešitve izkazujejo potrebe po določenih jezikovnih virih.

3.1 Metodologija

Zbirali smo objave uporabnikov na Facebooku v skupini *Prevajalci, na pomoč!*, v kateri si prevajalci pomagajo pri jezikovnih težavah. Objave smo zbirali ročno med novembrom 2014 in januarjem 2015, naš cilj pa je bil zbrati 100 relevantnih objav. Od 223 zbranih najaktualnejših objav smo izločili vse tiste, ki za našo raziskavo niso relevantne (npr. zahteve po prevodih v tuje jezike, ponudbe za delo, razvedrilne objave in ostale objave, ki niso konkretno povezave z jezikom). Nekatere od 100 zbranih relevantnih objav so vsebovale več vprašanj, zato smo jih razdelili in dobili skupno 106 vprašanj. Vprašanja smo nato ročno označili glede na kategorije, ki so bile določene na podlagi gradiva samega.

V nadaljevanju je predstavljen nabor kategorij, ki predstavljajo različne scenarije težav, s katerimi se prevajalci soočajo pri delu.

Scenarij 1: Ne vem ničesar, potrebujem prevod. → Prevajalec o izrazu v tujem jeziku ne ve ničesar in prosi za pomoč tako pri razumevanju kot pri iskanju izraza v slovenščini.

[1] Untreated redwood, kateri les je to pri nas (če sploh je pri nas) oz. kako se prevaja?

[2] Kako prevesti »digital by default«?

Scenarij 2: Poznam koncept, a ne termina/izraza. → Prevajalec pozna koncept, ki ga označuje izraz v tujem jeziku, ne pozna pa slovenskega izraza zanj – najpogosteje gre v takšnem primeru za termin oziroma izraz, ki v slovenščini še nima ustaljene oblike.

[3] Kako prevajamo »rib tips«? http://en.wikipedia.org/wiki/Pork_ribs

[4] Ko na začetku sabljanja trener reče: Fence. Fencers, ready? Fencers, salute. Fencers at the ready. - Začetni položaj? Single lunge.

Scenarij 3: Iščem prevod, imam rešitev – je prava? → Prevajalec je že prišel do rešitve, a ni prepričan v njeno pravilnost. Potrebuje potrditev.

[5] heritage sites - dediščinske točke (??) site of interest (muzeji, galerije, gradovi, jame itd.) - točke interesa (??)

[6] Zanima me, kako prevesti »Province of Treviso« in »Liguria Region« v slovenščino? V virih, ki sem jih našla, je Treviso pokrajina, Ligurija pa dežela ali regija.

Scenarij 4: Izraz bom določil sam – potrebujem ideje. → Prevajalec se je že odločil, da bo izraz v slovenščini določil sam – najpogosteje zato, ker ta še ni ustaljen ali predpisan v virih. Pogosto želi biti tudi ustvarjalen (npr. pri podnapisih, sinhronizaciji in v leposlovju).

[7] kako prevajate »performance« v kontekstu ličil/kozmetike? v mojem primeru govori o »...impeccable performance of make-up products...«. Obstojnost, učinkovitost? Kaj bi se uporabilo v slo. na tem mestu?

[8] »Showcase Festival« - v Slo -any ideas?

Scenarij 5: Imam dve varianti ali več – katera je ustreznejša? → Prevajalec ima na voljo več rešitev, a ni prepričan, katera je pravilnejša oziroma pomensko ustreznejša. Potrebuje potrditev.

[9] Se »skeeball« slučajno že sloveni kot recimo skibol (kot bejzbol)? Sicer močno dvomim, a vseeno vprašam. Hvala.

[10] Kako se operira z besedo 'quadrennial', torej dogodek/festival, ki se odvija vsake štiri leta? govorim konkretno o dogodku, ki se imenuje v ang. Prague Quadrennial, v češčini Pražský Quadrennial (ampak čehi mam občutek raje puščajo v nekih mednarodnih varjantah), kako bi bilo potem v slovenščini, Praški kvadrienal ali kar Praški Quadrennial?

Scenarij 6: Moja rešitev ni ustrezna – rad bi alternativo. → Prevajalec je prišel do rešitve, a se zaveda, da ni ustrezna. Išče alternativo.

[11] Takole iz firbca me zanima, ali kdo ve, če se je za »think tank« že našla boljša rešitev od precej neposrečenega, čeprav dokaj uveljavljenega »možganskega trusta«.

[12] A mogoče obstaja kak prevod za push up modrček?

Scenarij 7: Kaj pomeni ta izraz? → Prevajalec ne razume, kaj pomeni izraz v slovenščini. Potrebuje razlago.

[13] a kdo ve, kaj pomeni »tramak«? Kontekst: se omejenost skladiščnih površin rešuje s tramaki blaga.

[14] Bi kdo mogoče vedel, kaj v obrazcu pričakujejo pod 'geneza priimka'?

Scenarij 8: Kaj menite o lektorjevi rešitvi? → Prevajalec se ne strinja z lektorjevim popravkom in išče bodisi pojasnilo, zakaj se je lektor tako odločil, bodisi potrditev, da je lektorjev popravek nelogičen.

[15] LOL, lektorca me je ubila, ko je povedala, da SP predlaga odtisoček. Kaj pa menite vi, vesoljno strokovno občéstvo?

[16] Pišem o bodybuildingu in imam bodybuilding tako pogosto, da mi že skozi ušesa ven gleda. V skupini ljubiteljev slovenščine je prišel en odgovor, da se reče bodybuildingu »atletska gimnastika«. (WTF?) Jaz bi raje uporabljala kar slovenjeno bodibilding in bodibilder.

Scenarij 9: Potrebujem izraz v standardni slovenščini. → Prevajalec za nek koncept pozna izraz v nestandardni slovenščini, a ga v prevodu zaradi omejitev registra ne more uporabiti. Potrebuje izraz v standardni slovenščini.

[17] Mogoče kdo ima babico ali dedija zraven, ki ve kako se reče taki vrsti peči (srb. kraljica peći)? [slika]

[18] mogoce kdo ve, kako po slovensko recemo prskalici? [slika] Zaenkrat sem nasla samo »carobna svecka«, ma me ne preprica.

Scenarij 10: Zanima me etimologija besede. → Prevajalca zanima izvor določene besede.

[19] ima morda kdo pri roki etimološki slovar ali informacijo o izvoru besede KMET?

3.2 Rezultati

V spodnji tabeli so zbrana vprašanja razvrščena po scenarijih.

Tabela 1: Delež posameznih scenarijev.

Scenarij	Število vprašanj	Odstotek vprašanj
Scenarij 1: Ne vem ničesar, potrebujem prevod.	20	19
Scenarij 2: Poznam koncept, a ne termina/izraza.	33	31
Scenarij 3: Iščem prevod, imam rešitev – je prava?	14	12
Scenarij 4: Izraz bom določil sam – potrebujem ideje.	11	10
Scenarij 5: Imam dve varianti ali več – katera je ustreznejša?	10	9
Scenarij 6: Moja rešitev ni ustrezna – rad bi alternativo.	9	9

Scenarij	Število vprašanj	Odstotek vprašanj
Scenarij 7: Kaj pomeni ta izraz?	4	4
Scenarij 8: Kaj menite o lektorjevi rešitvi?	2	2
Scenarij 9: Potrebujem izraz v standardni slovenščini.	2	2
Scenarij 10: Zanima me etimologija besede.	1	1
Skupno	**106**	**100**

Rezultati kažejo, da se prevajalci najpogosteje srečujejo z jezikovnimi zagatami, ki jih bodisi sploh ne razumejo bodisi jih ne znajo izraziti.

Potrebam iz scenarija 1 (*Ne razumem ničesar, potrebujem prevod*) je z enojezičnim slovarjem slovenščine nemogoče zadostiti, saj gre za vprašanja o izrazih v tujem jeziku, pri katerih bi prevajalec potreboval dvojezični slovar (pogosto specializiranega).

Pri ostalih scenarijih, ki so bolj usmerjeni v slovenski del prevajalskega procesa, pa bi slovar lahko bil v pomoč, saj gre pogosto za primerjavo več rešitev.

Scenarij 2 (*Poznam koncept, a ne termina/izraza*) je problematičen predvsem zato, ker prevajalec najpogosteje išče zelo specializiran izraz s specifičnega področja (npr. avtomobilizem, strojništvo), zato je malo verjetno, da bo tak izraz vključen v slovar, ki vsebuje pretežno splošno slovenščino. Slovar pa bo v vsakem primeru vključeval terminologijo, zato je oznaka specifičnega področja pomemben del slovarskega opisa, prav tako tudi stopnja terminologizacije, torej podatek o tem, ali gre za termin, ki se uporablja zgolj v stroki ali pa je že prešel v rabo tudi v drugih, manj specializiranih besedilih.

Pri scenariju 3 (*Iščem prevod, imam rešitev – je prava?*) prevajalec rešitev že ima, a dvomi v njeno pravilnost, zato išče bodisi potrditev, da rešitev lahko uporabi, bodisi popravek. Vzroki za dvom so različni: prevajalec morda ni prepričan, ali se neka stvar uporablja, ali pa ni prepričan, da je pomen res takšen, kot si ga predstavlja. V takšnih primerih bi bilo ključno, da slovar vključuje večje število korpusnih zgledov ali pa da je slovarska informacija povezljiva s korpusi (in s tem z vrsto različnih kontekstualizacij).

Pri scenarijih 4 (*Izraz bom določil sam – potrebujem ideje*), 5 (*Imam dve varianti ali več – katera je ustreznejša?*) in 6 (*Moja rešitev ni ustrezna – rad bi alternativo*) je smiselno izpostaviti sinonimijo v slovarju. V vseh treh scenarijih bi bila namreč prevajalcu lahko v pomoč možnost primerjave sinonimov ali sorodnih izrazov v slovarju, bodisi po pogostosti rabe, po kvalifikatorjih, po kolokacijah, po zgledih rabe ipd.

Scenarij 7 (*Kaj pomeni ta izraz?*) je od vseh najbolj preprost, saj ga razreši ustrezen pomenski opis iztočnice. Pomembno je, da se v slovarju relevantne pomenske informacijo hitro najdejo (npr. s pomočjo pomenskega menija) in da jih uporabnik lahko uspešno razbere. Primeri iz tega scenarija pogosto vključujejo iztočnice, ki so novejšega izvora ali pa še niso povsem uveljavljene v jezikovni rabi. Razmisliti bi bilo treba o možnostih posodabljanja slovarja in dodajanja novih iztočnic glede na trende v jezikovni rabi, saj ta scenarij v veliki meri kaže na vrzel med hitrim razvojem jezika in tradicionalno statičnostjo jezikovnih priročnikov.

Pri scenariju 8 (*Kaj menite o lektorjevi rešitvi?*) bi si prevajalec lahko pomagal s kvalifikatorji in informacijami o pogostosti rabe določene iztočnice, kar znova kaže na potrebo po povezovanju slovarja z drugimi viri (npr. s korpusi). Pomembno je tudi, da ima slovarski uporabnik dostop do čim več referenčnih virov, s katerimi lahko potrdi ali ovrže svojo odločitev.

Scenarij 9 (*Potrebujem izraz v standardni slovenščini*) je nekoliko problematičen, saj kaže na razhajanje med standardno in nestandardno slovenščino. Treba bi bilo torej razmisliti, na kakšen način in v kolikšni meri bi v slovar lahko dodali tudi iztočnice iz nestandardne slovenščine, kar odpira številna nova vprašanja s področja lematizacije in zapolnjevanja leksikalnih vrzeli med standardom in nestandardom.

Scenarij 10 (*Zanima me etimologija besede*) se je v tej pilotni raziskavi izkazal za zelo redkega, a tudi kaže na potrebo po tem, da se novi slovar povezuje z drugimi jezikovnimi viri (v tem primeru s kakovostnimi etimološkimi priročniki). Skladno z idejo novega slovarja, da je uporabniku čimbolj prijazen, pa bi se lahko tovrstno informacijo v poenostavljeni obliki vključilo tudi v geselske članke.

3.2.1 Kako bi slovar lahko bil v pomoč?

S tem parametrom smo pri vsakem vprašanju označili, kateri vidik slovarja bi bil prevajalcu morda v pomoč pri razreševanju težav (v primerih, ko slovar ne bi bil v pomoč, smo ta parameter izpustili). V tabeli so predstavljene vrednosti parametrov in njihova pogostost.

Tabela 2: Delež vsebinskih slovarskih sklopov pri rešitvi težave.

Kako bi slovar lahko bil v pomoč	Število pojavitev	Odstotek pojavitev
Terminologija	35	32
Sinonimi	27	24
Pomenski opis	25	22
Kvalifikatorji	9	8
Besedne družine	7	6
Razlaga pomena	3	3
Primeri rabe	2	2
Etimologija	1	1
Idiomatika	1	1
Pravopisna pravila	1	1
Skupno	**111**	**100**

Opazimo lahko, da so pravzaprav vse informacije relevantne. Zanimivo je, da bi bili primeri rabe le redko v pomoč, saj gre pri vprašanjih, ki jih zastavljajo prevajalci, pogosto za zelo specifične kontekste, ki jih primeri rabe v slovarju ne razrešujejo. Poleg samega pomenskega opisa iztočnice v slovarju pa sta za prevajalce izrazito pomembna vidika še terminologija in sinonimija.

Terminologija odpira načelno vprašanje o količini specializirane leksike v slovarju, pri čemer je z vidika prevajalcev razvidno, da je to ena od ključnih informacij, po kateri se povprašuje. Vprašanje seveda je, ali tovrstna vprašanja res lahko razreši enojezični splošni slovar ali pa gre za tako specializirano leksiko, da bo na tovrstne zadrege lahko odgovoril šele terminološki vir.

Prav gotovo pa enojezični slovar lahko razrešuje vprašanje sinonimije. Tu se poleg same predstavitve sinonimnih možnosti v slovarju odpira vprašanje razmerja do tezavra oz. vprašanje, kako tovrstne informacije združevati v enem jezikovnem viru, da bodo za uporabnika hitro in intuitivno dostopne.[3]

3.2.2 Način reševanja

Pri tem parametru smo opazovali, kdo vse je udeležen pri reševanju jezikovne težave in kateri viri (če sploh) so pri tem uporabljeni.

3 Veliko tujih enojezičnih slovarjev, zlasti angleških (npr. Oxford Dictionary of English, Merriam-Webster, Collins, Macmillan) na spletnih straneh ponuja tako enojezični slovar kot tezaver, in sicer tako eksplicitno (oba sta navedena med viri) kot implicitno (uporabnik izvede iskanje, med zadetki pa je del, namenjen sinonimom).

Tabela 3: Delež virov pri reševanju težave.

Način reševanja	Število pojavitev	Odstotek pojavitev
Prevajalci	96	63
Viri	30	20
Brskanje po spletu	19	13
Strokovnjaki	6	4
Skupno	**151**	**100**

Ker gre za skupnost prevajalcev, je logično, da so v proces razreševanja določene težave vpeti prevajalci sami, prav tako pa je pomemben vir informacij splet na sploh. V nadaljevanju smo zato želeli natančneje določiti, kateri so bili tisti eksplicitno navedeni viri, ki so jih prevajalci uporabili pri reševanju jezikovnih težav[4] (večina se pojavi redko, zato odstotki tu niso našteti).

Tabela 4: Vrsta uporabljenih virov pri reševanju težave.

Uporabljeni viri	Število pojavitev
Spletne strani	30
Google	9
SSKJ	8
Strokovnjaki	6
Wikipedija	6
Diplomska in magistrska dela	3
Spletni članki	2
Eur-Lex	2
Termania	2
Gigafida	1
Nova beseda	1
Pravni slovar	1
Slovenski pravopis	1
Avtomobilski slovar	1
Angleško-slovenski slovar Oxford DZS	1
Etimološki slovar (Snoj)	1
Evroterm	1
iSlovar	1
Seznam tujih zemljepisnih imen v slovenskem jeziku	1
Študentska skripta	1
Terminologišče	1
Veliki leksikon DZS	1

4 Vključili smo zgolj vire, ki so jih prevajalci omenjali ali predlagali v svojih razpravah. Po vsej verjetnosti so prevajalci uporabili tudi številne druge vire, zato pregled virov v tem prispevku ni celosten.

Gre za zelo razpršene vire, med katerimi izstopajo spletne strani, za katere prevajalci menijo, da zaradi specializiranosti leksike lahko podajo ustrezno rešitev. Tudi tu se izkaže, da je splošni spletni brskalnik (v našem primeru Google) tisti, ki je velikokrat izhodišče iskanja rešitev, kar dobro ilustrira način pristopanja k reševanju problema, kjer se v prvem koraku ne izbere specializiranega vira, temveč informacijo na spletu na sploh.[5]

Zanimivo je, da je kljub sorazmerno velikemu številu vprašanj, ki se razrešujejo glede na kontekst, uporaba korpusnih virov, kjer bi vpogled v kontekstualizacijo lahko privedel do ustrezne rešitve, praktično zanemarljiva. To lahko pomeni, da se prevajalci pri svojem delu ne zanašajo na korpusne vire ali pa da z njimi niso dobro seznanjeni.

3.2.3 Odnos spraševalca

Pri tem parametru smo opazovali, ali je spraševalčevo vprašanje čustveno zaznamovano oziroma ali se v vprašanju kaže njegov odnos do problema. Rezultati so predstavljeni v spodnji preglednici.

Tabela 5: Deleži izraženega osebnega odnosa pri zastavljanju vprašanja.

Odnos spraševalca	Število pojavitev	Odstotek pojavitev
Neizražen	75	71
Dvom	23	21
Kreativen	4	4
Negativen	2	2
Nelogičnost	1	1
Pokroviteljski	1	1
Skupno	**106**	**100**

Kot je razvidno iz preglednice, je v večini primerov odnos spraševalca neizražen oziroma nevtralen, kar kaže na dejstvo, da ta skupina uporabnikov svoje jezikovne dileme največkrat izraža čustveno nezaznamovano. Iz tega lahko sklepamo, da prevajalci kot poklicni jezikoslovci k razreševanju problemov pristopajo z racionalnimi pričakovanji. V 21 odstotkih primerov je prisoten dvom, najpogosteje takrat, ko prevajalec že ima rešitev, a v njeno pravilnost ni prepričan. Občutki negativnosti, nelogičnosti in pokroviteljstva se pokažejo le v marginalnih primerih.

5 Da je to splošen trend, lahko sklepamo tudi po rezultatih sorodnih raziskav. Lorentzen in Theilgaard (2012) sta npr. ugotovila, da 49 % uporabnikov do slovarja *Den danske ordbog* dostopa prek iskalnika, 33 % neposredno in 18 % prek drugih spletnih strani. Ko sta izboljšala indeksiranje slovarja, se je obisk prek iskalnika dvignil na 84 %.

V 4 odstotkih primerov se je pokazal ustvarjalen odnos, pri čemer prevajalec najpogosteje išče domiselne, hudomušne ali »lepo zveneče« predloge za svoj prevod.

3.3 Diskusija

S pilotno raziskavo smo želeli ugotoviti, v kolikšni meri lahko z inovativnim pristopom, ki analizira verbalizirane prevajalske težave znotraj samoorganizirane prevajalske skupnosti, dobimo podatke o vrsti informacije v enojezičnih (slovarskih) virih, relevantnih za prevajalce. Zavedamo se, da je raziskavo treba obravnavati kot pilotno, saj je bila kot takšna tudi načrtovana, in da rezultatov ne moremo posploševati na celotno prevajalsko skupnost, temveč nam lahko služijo le kot izhodišče za nadaljnje raziskave.

Ob večini predstavljenih scenarijev se ponujajo predlogi, katere informacije bi bilo treba vključiti v slovar, da bi bile prevajalcu v pomoč. Čeprav se pri reševanju izpostavljenih prevajalskih problemov pojavlja vrsta informacij, ki so za prevajalce v okviru leksikografskih opisov relevantni, na podlagi naše analize lahko izpostavimo dve ključni vprašanji: vključevanje in obravnava (a) specializirane leksike in (b) sinonimije. Pri prvem gre za načelno odločitev o obsegu specializirane leksike v slovarju, vprašanje pa je, ali tovrstna vprašanja res lahko razreši enojezični splošni slovar ali le specializirani. Pri vprašanju sinonimije pa je razvidno, da prevajalci v slovarskem opisu pričakujejo združevanje informacij klasičnega slovarja in tezavra. Rezultati kažejo, da bi že ta dva vidika, če bi bila v novem slovarju ustrezno in zadostno zastavljena, razrešila precejšen delež zaznanih prevajalskih težav (oziroma bi vsaj v veliki meri pripomogla k njihovemu razreševanju).

Način iskanja po virih tudi v našem primeru kaže na to, da je k predstavitvi slovarskih informacij treba pristopati na povsem nov način. Klasična slovarska predstavitev za digitalni medij preprosto ni učinkovita, saj je prvi impulz prevajalcev iskanje na način, ki so ga oblikovali spletni iskalniki. Obenem je ključno tudi, da se razmisli o načinih, kako zagotoviti, da bo novi slovar čim bolj uspešno dohajal realno jezikovno rabo in bo vseboval tudi novejše iztočnice, s katerimi se prevajalci srečujejo pri svojem delu.

4 ZAKLJUČEK

Med uporabniškimi raziskavami slovarjev so ena od precej jasno definiranih uporabniških skupin prevajalci. Morda je bila prav zaradi profiliranosti skupine v tujini opravljena vrsta raziskav o uporabi slovarjev pri prevajalcih in o

njihovih pričakovanjih glede slovarskih informacij. Navezale so se na predhodne študije uporabe slovarjev za učenje angleščine kot tujega jezika, iz katerih so izšle tudi raziskave uporabe slovarjev pri izobraževanju prevajalcev. Za slovenski prostor je le malo analiz o uporabi slovarjev pri posameznih skupinah uporabnikov, med njimi pa je vendarle nekaj takih, ki se ukvarjajo z uporabo slovarjev pri študentih prevajanja oz. pričakovanji študentov prevajalstva glede oblike in vsebine slovarskega opisa.

V članku smo predstavili enega od možnih pristopov pri analizi uporabniških potreb prevajalcev: na podlagi diskurza prevajalske skupnosti na omrežju Facebook v skupini *Prevajalci, na pomoč!* smo zbrali vprašanja članov in z njihovo pomočjo oblikovali različne scenarije, ki na podlagi vprašanja predvidevajo njegovo razrešitev. Metoda se je v tem okviru pokazala kot uspešna, saj smo na ta način sorazmerno dobro opisali uporabniške potrebe, hkrati pa tudi vsebinske sklope težav ter poti njihovih reševanj.

Z raziskavo se je pokazalo, da je s predlagano metodo mogoče opisati tipične scenarije jezikovnih težav, s katerimi se prevajalci soočajo pri delu, na ta način pa je mogoče priti tudi do podatkov o vsebinskih sklopih predstavljenih problemov. Z analizo razreševanja problemov pridemo tudi do relevantnih informacij o uporabljenih virih za razrešitev določenega problema.

Za celostni vpogled v potrebe prevajalske skupnosti bo v prihodnje treba opraviti dodatne raziskave, tudi z drugačnimi metodološkimi pristopi, od klasičnega anketiranja in intervjujev do študij, v katerih bomo v tej pilotni raziskavi opisane scenarije bolj ciljno testirali v realnih prevajalskih okoljih. Le na ta način bomo prišli do kvalitetnih podatkov o uporabnikih in njihovih pričakovanjih, tudi glede pričakovanj jezikovnega opisa v enojezičnem slovarju slovenskega jezika.

Slovarski uporabniki – ustvarjalci: ustvarjati v jeziku in z jezikom

Vesna Mikolič

Abstract
This paper presents a pilot study conducted among language guide users who work in creative jobs characterised by the creative use of language, such as writers, scientists, journalists and copywriters. We were interested in how their language awareness is manifested through their use of language guides. The survey showed that every group uses both printed and electronic language guides from time to time, but the frequency of use varies from group to group and from generation to generation.

Keywords: language guides, dictionary, user survey, language awareness, creativity

Ključne besede: jezikovni priročniki, slovar, uporabniška raziskava, jezikovna zavest, ustvarjalnost

1 UVOD

Uporabniki jezikovnih priročnikov, ki slednje uporabljajo za poklicne namene, so zelo široka in raznolika skupina. V našo ožjo raziskavo smo zajeli tiste poklicne profile, za katere je značilna (redna) produkcija besedil, ki je namenjena širši javni rabi in je do določene mere odraz avtorske ustvarjalnosti. Diskurzi literarnih ustvarjalcev, znanstvenikov, oglaševalcev in novinarjev, ki smo jih upoštevali v raziskavi, se med sabo po namenu svojega jezikovnega izražanja sicer razlikujejo, kljub temu je prav področje ustvarjalnosti tisto, ki je vsem štirim kategorijam skupno.[1]

2 SPOROČANJSKI NAMENI USTVARJALCEV

Pri analizi namena njihovega ustvarjanja si lahko pomagamo s teorijo govornih dejanj in delitvijo človekove govorne dejavnosti na osnovi štirih osnovnih vplivanjskih vlog, tj. spoznavne, sporazumevalne, izvršilne in umetnostnoizrazne (Mikolič 2007; Skubic 1994/1995; 2005). Namen literarnega ustvarjanja je umetnostnoizrazni, literarni ustvarjalec in ustvarjalka s sebi lastnim estetskim izrazom izpovedujeta subjektivni pogled na stvarnost. Bolj kot je njegov/njen svet enkraten in večplasten, bolj dragocena je njegova/njena literatura, seveda pod pogojem, da je sama v sebi koherentno povezana in zato prepričljiva. Bralec avtorja zanima šele v drugi fazi, tedaj ko želi, da literarno delo zaživi svoje lastno življenje. Tudi znanstvenika/znanstvenico k raziskavam primarno ne žene želja po komuniciranju spoznanj publiki, pač pa ustvarjalen odnos do obstoječe stvarnosti, v kateri vidi ves čas nove in nove izzive, še neraziskana področja. Prav tako je torej tudi osnova za znanstveni diskurz povsem subjektivna in vsebuje veliko mero ustvarjalnega mišljenja, saj morata biti znanstvenica in znanstvenik sposobna pogledati na že znana dejstva drugače, z novega, morda doslej neznanega vidika. Vendar pa za razliko od literarne umetnosti znanost sodi v spoznavno polje, saj je znanstvenikov prvi namen odkrivati še nepoznano stvarnost in razmerja v njej, širiti obstoječe in ustvarjati novo znanje, znanstveno delovanje ima torej primarno spoznavni namen. Sprva subjektivni uvid v zunajjezikovno dejanskost mora sicer znanstvenik skozi raziskavo potrditi z dokazi in ga s tem objektivizirati, kljub temu mu ostaja ustvarjalnost metod, možnost povezovanja nepovezanega in razdruževanja navidez nezdružljivega ter nenazadnje vedno novi in novi ustvarjalni vidiki in vpogledi. Publicistični in oglaševalski diskurz sodita v okvir sporazumevalnih dejavnosti v ožjem pomenu besede, saj je njuna osnovna značilnost orientiranost na naslovnika, njun primarni namen pa sporazumeti se z naslovnikom, posredovati mu sporočilo. Vendar pa naj bi pri tem v novinarskem

1 Zavedamo se, da je tudi prevajanje ustvarjalna dejavnost, a prevajalce na tem mestu izpuščamo, saj se s to uporabniško skupino ukvarja druga razprava v tej monografiji (Čibej et al. 2015).

govoru v glavnem prevladovala težnja po informiranju naslovnika, oglaševalski pa želi prejemnika predvsem prepričati o pozitivnih valencah besedilne reference. Pri tem tako novinarji kot oglaševalci iščejo inovativne, učinkovite komunikacijske strategije, še posebej to velja za oglaševanje, kjer je danes kreativnost sploh osrednji element in vsi ostali elementi uspešnega oglaševanja izhajajo iz njega (Jewler in Drewniany 2005).

Kot smo videli, se po obsegu in metodah ustvarjalnosti predstavljene kategorije diskurzov med sabo razlikujejo. Za razliko od literarnega dela, ki je torej tem bolj cenjeno, čim bolj je avtor koherenten v svojem subjektivnem svetu in lastnem izrazu, pa mora znanstvenik za svojo verodostojnost svoj začetni inovativni, subjektivni pogled na problematizirano stvarnost v teku raziskave s pomočjo dokazov objektivizirati. Podobno razliko bi lahko našli med novinarji in oglaševalci, saj imajo slednji skladno z namenom oglaševanja nekoliko več ustvarjalne svobode, čeprav profesionalna etika narekuje zavezanost resnici neke objektivne danosti. Kljub temu pa lahko za vse štiri kategorije avtorjev besedil rečemo, da so pri svoji jezikovni produkciji vsaj do določene mere subjektivni in ustvarjalni; besedila, ki nastajajo na osnovi tega ustvarjalnega pogleda, pa so – slej ko prej – namenjena širši publiki. Na osnovi tega lahko predvidimo njihovo posebno občutljivost za jezik; v našem prispevku nas bo zanimalo, kako se ta jezikovna zavest kaže skozi njihove potrebe po slovarskih in drugih jezikovnih priročnikih.

3 CILJI IN METODE PILOTNE RAZISKAVE

Z vzponom družbenega jezikoslovja v šestdesetih let prejšnjega stoletja dalje so se uveljavila spoznanja, da je jezikovno sporočanje vedno interaktivno, da je vedno namenjeno naslovniku, aktualnemu ali potencialnemu (Schiffrin 1988), tako se je nekje istočasno začel tudi razvoj uporabniških raziskav, ki jih je zanimal tudi uporabnik jezikovnih priročnikov (glej uvodni prispevek Arhar Holdt). To je seveda razumljivo, saj so priročniki, tudi jezikovni, besedilna vrsta, ki je eksplicitno in primarno namenjena uporabniku. Preseneča pa, da v slovenskem prostoru tej temi doslej nismo posvečali dovolj pozornosti, razen nekaterih, ki so v zadnjih letih začeli opozarjati na pomen raziskav slovarskih uporabnikov (Logar 2009; Stabej 2009). Nedvomno se je potreba po spremembi fokusa, torej po usmerjenosti na uporabnika, pokazala tudi v nekaterih novejših projektih, ki jih zanimajo uporabniki jezika širše in so se odločili njihove potrebe načrtno spremljati; tako so npr. raziskovalci projekta Sporazumevanje v slovenskem jeziku želeli ugotoviti, »na katerih mestih imajo pišoči pri pisanju v slovenščini **v resnici** težave«, na osnovi česar so bili v okviru projekta izdelani Slogovni priročnik in še številni drugi spletni jezikovni viri in orodja, naravnani na potrebe jezikovnih uporabnikov (gl. SSJ). K jezikovnim uporabnikom so usmerjene tudi različne spletne

svetovalnice, npr. na spletnih portalih Fran, ŠUSS. Zaradi hitrega razvoja komunikacijskih tehnologij in korenitih sprememb narave sporazumevanja v sodobni družbi pa je predvsem potrebno vzpostaviti redno spremljanje razvoja uporabniških potreb.

V pričujočem prispevku zato predstavljamo pilotno raziskavo med uporabniki jezikovnih priročnikov, ki delujejo v ustvarjalnih poklicih. Raziskava pomeni osnovo za nadaljnje preučevanje na tem področju in redno spremljanje načina dela tovrstnih uporabnikov jezika in njihovih potreb v zvezi z jezikovnimi priročniki.

Izhodiščno vprašanje je bilo, kako se jezikovna zavest posameznikov, ki se v svojih poklicih ustvarjalno ukvarjajo z jezikom, kaže skozi njihove potrebe po slovarskih in drugih jezikovnih priročnikih.

V ta namen so bili spomladi 2015 izvedeni polstrukturirani intervjuji (polovica v živo, polovica po e-pošti) s 30 ustvarjalci. V obliki polstrukturiranega intervjuja smo jim zastavili naslednja vprašanja:

1. Ali pri svojem delu uporabljate jezikovne priročnike?

   1. 1. Če da, katere?

2. S kakšnim namenom oziroma ob katerih vprašanjih posežete po njih?
3. Ali poznate spletne jezikovne priročnike oziroma orodja? (slovarji, korpusi …)

   3. 1. Če da, katere?

4. Ste imeli kakšen jezikovni problem, vprašanje, na katerega vam jezikovni priročniki niso dali odgovora?
5. Kaj pričakujete od slovarja slovenskega jezika? Kakšna mora biti po vašem njegova vsebina?

Informantje so tako v intervjujih v živo kot v tistih po e-pošti prosto odgovarjali na vprašanja, pri čemer so nekateri bolj, drugi manj uporabljali jezikoslovno terminologijo, povezano z rabo slovarjev in drugih jezikovnih priročnikov. V analizi sledimo njihovim formulacijam, obenem pa smo njihove odgovore smiselno povzeli v jezikoslovne kategorije in jih tudi prikazali v tabelaričnem in slikovnem gradivu.

Od 30 intervjuvanih ustvarjalcev je bilo deset literarnih ustvarjalcev, deset znanstvenikov, pet kreativcev/oglaševalcev in pet novinarjev različnih generacij (mlajša od 20-35 let, srednja od 35-50, starejša nad 50), spola in regionalne pripadnosti (iz Kopra oz. Pirana, Ljubljane in Maribora). Med literarnimi ustvarjalci so bili pesniki, prozaisti in dramatiki (3 ženske, 7 moških), intervjuvani znanstveniki so bili s področja humanistike (brez jezikoslovcev), družboslovja in naravoslovja

(6 žensk, 4 moški), od petih novinarjev sta bili dve sodelavki tiskanih medijev, sodelavka in sodelavec radijskega ter sodelavka televizijskega medija, med kreativci so bili sodelavka oglaševalske agencije, oblikovalka s statusom svobodne umetnice, sodelavec in sodelavka PR-službe večjega podjetja oziroma ustanove in upokojenka, prej zaposlena v PR-službi večjega podjetja. Glede na majhen vzorec imajo seveda ugotovitve omejeno vrednost, kljub temu pa so pokazale na nekatere zanimive poteze jezikovne zavesti omenjenih profilov ustvarjalnih piscev in na nekatere njihove podobnosti in razlike, kar vse bo potrebno, kot rečeno, še naprej in še bolj poglobljeno spremljati.

4 ODNOS LITERARNIH USTVARJALCEV DO JEZIKOVNIH PRIROČNIKOV

Zanimivo je, da smo pri tem profilu uporabnikov hitro evidentirali dve skrajnosti; namreč, na eni strani so redni uporabniki jezikovnih priročnikov, na drugi pa so tisti, ki se raje zanesejo na svoj jezikovni čut in jezikovnih priročnikov sploh ne uporabljajo ali pa le izjemoma.

Kaže, da je to delno generacijsko pogojeno. Uporabniki iz starejše in nekateri uporabniki srednje generacije redno uporabljajo tiskano izdajo Slovenskega pravopisa (SP), Slovarjev slovenskega knjižnega jezika (SSKJ), slovarjev tujk, ena intervjuvanka iz te skupine uporablja tudi slikovne slovarje, en informant pa občasno odpre tudi Toporišičevo slovnico ter Bezlajev ali Snojev etimološki slovar. Spletnih jezikovnih priročnikov in virov ne poznajo, kvečjemu pobrskajo po spletu s pomočjo spletnih iskalnikov, če jih zanima kakšna specifična jezikovna raba.

Drugi predstavniki srednje in predstavniki mlajše generacije uporabljajo jezikovne priročnike zelo redko, od teh tiskane priročnike Verbinčev Slovar tujk, Oxfordov slovar angleškega jezika in spletne priročnike SSKJ, SP, Wikipedijo. Drugih spletnih priročnikov in virov (npr. korpusov) ne poznajo.

Po jezikovnih priročnikih posegajo pri pisateljevanju, prevajanju, ena avtorica navaja tudi potrebe samoizobraževanja. Jezikovne priročnike uporabljajo za iskanje predvsem pravopisnih in slovničnih informacij, včasih natančnega pomena ali pri rabi oz. tvorbi neologizmov (zanima jih, ali beseda že obstaja v slovarju), poizvedujejo za besedotvornimi značilnostmi, zato da odkrijejo jezikovne zakonitosti pri tvorbi neologizmov, zanimajo pa jih tudi stilno zaznamovane besede, večpomenskost, pri ritmiziranih besedilih tudi naglas.

Sicer se v glavnem ne spomnijo, da bi imeli kakšno konkretno jezikovno vprašanje, na katerega jim jezikovni priročniki niso dali odgovora, nekateri pa navajajo,

da se včasih s kakšno pravopisno rešitvijo niso strinjali ali da so v slovarju pogrešili kakšno besedo; en avtor pa izpostavi, da v slovarju pogosto pogreša kakšne morda danes tudi nestandardne besede, ki pa so se ohranile v narečjih in so tudi del slovenskega jezika.

Bolj zgovorna so njihova pričakovanja glede slovarja slovenskega jezika, ki jih lahko strnemo v nekaj lastnosti: slovar naj bo predvsem pregleden, poleg tega tudi čim bolj izčrpen, opisi in primeri rabe čim obširnejši, izhajajo naj iz živega govora, dostopen naj bo tako v knjižni obliki kakor na spletu. Sicer pa njihova pričakovanja lepo opišeta tudi spodnji izjavi:

> V slovarjih na primer pogrešam mnoge besede, ki so v vseh pogledih čisto slovenske besede, ki pa so se morda ohranile le v tem ali onem narečju. Posebej me zanima »jezikovna arheologija«, torej iskanje v jeziku skritih arhaičnih ostalin, tudi dragocenosti – gre na primer za jezikovni značaj, duhovno podstat tega značaja, tega »duha« jezika, ki ga je mogoče zaslediti v korenih in ostalih fonemih, tudi v sintaksi itn. Tem komponentam se jezikoslovci ali etimologi najraje izognejo, kar je morda z znanstvenega stališča povsem opravičljivo – medtem ko pisatelji pogosto iščemo prav to »magično« živost, saj je jezik živo izročilo, medij, skozi katerega pisatelj »priziva« duhovnost, duha tudi že zdavnaj mrtvih rodov, ki so ta jezik ustvarjali. O tem »duhu« jezikoslovje seveda v glavnem molči, kar pa ne more in ne sme pomeniti, da tega v jeziku ni. Po mojem gre pri tem celo za bistvo vsakega jezika in njegove ustvarjalne rabe, ki pa se jezikoslovju v glavnem izmika.
>
> (informant pisatelj iz Ljubljane, rojen l. 1958)

> /…/ da v slovarju ne bo umetno proizvedenih besed, ampak da se bo ravnal po ljudskem, lepem slovenskem jeziku in ne bo prirejal in izumljal besed. /…/ Slovar jemljem kot neko drugo mnenje, in ne kot absolutno resnico.
>
> (informant pesnik iz Pirana, rojen l. 1984)

5 ODNOS ZNANSTVENIKOV DO JEZIKOVNIH PRIROČNIKOV

Raziskovalci občasno vsi uporabljajo jezikovne priročnike. Včasih uporabijo tiskano izdajo SP in SSKJ, poznajo in uporabljajo pa več spletnih jezikovnih priročnikov, tako navajajo naslednje: spletni SSKJ, SP, PONS in druge dvojezične spletne slovarje, slovarje angleškega in drugih tujih jezikov (nemškega, italijanskega), slovarje klasičnih jezikov (stara grščina), terminološke slovarje (geografskega), večjezično terminološko zbirko Evroterm, prevajalnik Google Translate, Amebis Pressis, Besana, Termania, ne poznajo pa korpusov slovenskega ali tujih jezikov. Srednja in mlajša generacija znanstvenikov uporabljata skoraj izključno samo

spletne jezikovne priročnike, po nasvete in mnenja se obračajo tudi na spletne svetovalnice in forume.

Jezikovne priročnike potrebujejo včasih za iskanje pravopisnih in slovničnih informacij, iščejo predloge za sopomenke, alternativne izraze, večinoma pa so njihove potrebe povezane s tvorbo in rabo terminov. Tako pri prevajanju strokovne literature v slovenščino iščejo slovenske ustreznice tujim terminom (npr. angl. *aspiration economy* – ekonomija učinkovitosti), iščejo ustrezne korene za tvorbo novih terminov (npr. angl. *citizenisation* – državljenje), v primeru terminoloških dvojnic iščejo slovenska poimenovanja. Prav tako pa jezikovne priročnike, predvsem dvojezične slovarje ali slovarje tujih jezikov uporabljajo, ko pišejo v tujem jeziku, včasih iščejo prevodne ustreznice, včasih samo preverjajo zapis besed.

S prevajanjem so povezane tudi težave, ki jih niso zmogli rešiti s pomočjo jezikovnih priročnikov, in sicer včasih ne najdejo ustreznih prevodov, včasih pa ne najdejo zadostne razlage, da bi se odločili za nov termin. Predvsem jih motijo splošne, premalo natančne in premalo strokovne pomenske razlage, ki poleg tega pogosto ne upoštevajo večdisciplinarne rabe nekega izraza. Prav zato od slovarja slovenskega jezika pričakujejo, da je bogato opremljen s primeri rabe, da so primeri rabe tudi v kompleksnejših stavčnih strukturah ter v okviru različnih kontekstov in področij, da so omenjene različne posebnosti pomena in rabe. Tudi raziskovalci izpostavljajo pomen preglednosti, jasne strukture slovarja, obenem pa naj upošteva možnosti novih tehnologij. Njihova pričakovanja so dobro povzeta v spodnjih izjavah:

> Ne vem, nikoli nisem razmišljala na takšen način… Da je posodobljen, kar pomeni, da vključuje tudi sodobnejše besede, morda strokovno terminologijo, tudi tujke …
>
> (informantka znanstvenica s področja družboslovja iz Kopra, rojena l. 1971)

> Slovar naj predstavi različne kontekste besede, pomembna pa je tudi forma, jasna, pregledna, obenem naj bo slovar interaktiven, z vsemi možnostmi, ki jih omogočajo nove tehnologije.
>
> (informant znanstvenik s področja humanistike iz Ljubljane, rojen l. 1977)

> Pojem naj bo predstavljen interdisciplinarno, da se lahko uporabi tudi na širših področjih, saj je življenje širše, se ne omejuje na posamezne stroke. Za reševanje izzivov sodobnega časa je potrebno oblikovati interdisciplinarna področja (npr. za problematiko voda, lesa) in tem področjem mora slediti tudi ustrezno izrazje.
>
> (informantka znanstvenica s področja družboslovja iz Maribora, rojena l. 1967)

6 ODNOS NOVINARJEV DO JEZIKOVNIH PRIROČNIKOV

Tudi novinarji so dokaj redni uporabniki jezikovnih priročnikov v knjižni in spletni obliki, uporabljajo pa predvsem SSKJ in SP, poleg tega občasno uporabljajo še slogovne priročnike Janeza Gradišnika, dvojezične slovarje, različne pravniške priročnike, Krajevni leksikon, različne enciklopedije, včasih pobrskajo po Wikipediji in forumih, kjer se rešujejo jezikovne zagate, samo ena intervjuvanka iz te skupine brska tudi po korpusih (npr. Gigafida). Predstavniki mlajše generacije uporabljajo izključno spletne jezikovne priročnike in vire, čeprav sodobnejših spletnih jezikovnih portalov z zbranimi več priročniki in viri niti oni ne poznajo.

Jezikovne priročnike uporabljajo pri novinarskem delu, tj. pri pisanju člankov, zanimata jih predvsem pravopis in slovnica, npr. sklanjanje tujih imen in raba velike začetnice, pogosto po jezikovnih priročnikih posežejo, ko si želijo opredelitve razmerja med novejšim besedjem in jezikovno normo. Včasih jih zanima tudi terminologija, predvsem pravna, novinarje iz radijskega in televizijskega medija pa tudi pravilno naglaševanje. Ena od novinark je izpostavila uporabo jezikovnih priročnikov tudi v prostem času oziroma v družini (pri pomoči osnovnošolskemu otroku in študentki, lektoriranju diplomskih del).

Ob jezikovnih nejasnostih se pogosto obrnejo tudi na lektorja (kjer ga imajo) ali novinarske kolege. V slovarjih pogrešajo novejše besedje, pogosto namreč besed, ki jih iščejo, v slovarjih ne najdejo.

Od slovarja slovenskega jezika pričakujejo preglednost in ekonomičnost, saj zaradi narave svojega dela nimajo časa za dolgotrajno reševanje jezikovnega problema in morajo odgovor na svoje vprašanje dobiti zelo hitro. To potrebo izpričuje tudi spodnja izjava:

> Tempo mojega dela je izjemno hiter in včasih se preprosto nimam časa poglobiti v posamezno vprašanje in najti rešitev.
>
> (informantka novinarka iz Kopra, roj. 1987)

Poleg tega si želijo, naj bo slovar ažuren in naj torej spremlja razvoj novega besedišča, pri čemer naj bo odprt do novih izrazov, tudi tehnoloških, ki jih ponuja sodobna tehnologija. Predvsem pa naj ima pred očmi jezikovnega uporabnika, tako kot morajo biti na potrebe bralcev, poslušalcev in gledalcev naravnani tudi novinarji, o čemer priča naslednja izjava:

> Novinarji smo dolžni določena, tudi strokovna vprašanja približati bralcu in to lahko počnemo samo z razumljivim in slogovno ustreznim jezikom, zato si v slovarjih želimo preglednih pojasnil o pomenih in dejanski rabi besed.
>
> (informantka novinarka iz Kopra, roj. 1964)

7 ODNOS OGLAŠEVALCEV DO JEZIKOVNIH PRIROČNIKOV

Oglaševalci/kreativci so dokaj redni uporabniki jezikovnih priročnikov. Uporabniki srednje in starejše generacije redno uporabljajo tiskane izdaje SP, SSKJ, slovarjev tujk, Angleško-slovenskega slovarja, občasno Etimološkega slovarja, drugi uporabniki srednje in vsi uporabniki mlajše generacije pa uporabljajo iste priročnike na spletu, poleg teh pa še: slovarje spletnega portala BOS Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, Razvezani jezik – prosti slovar žive slovenščine, nekatere terminološke slovarje (npr. gledališkega), dvojezične spletne slovarje, slovarje angleškega in drugih tujih jezikov, Google Translate, ne uporabljajo pa besedilnih korpusov.

Jezikovne priročnike uporabljajo predvsem za iskanje pravopisnih in slovničnih informacij, poizvedovanje za natančnejšim pomenom, tudi stilno zaznamovanim, in za iskanje sopomenk zaradi izogibanja tujkam. Eden od intervjuvancev iz te skupine pa izpostavi, da se oglaševalci na drugi strani zavedajo trenda pretirane uporabe tujk v oglaševanju, v kar jih vodi želja po biti poseben, drugačen, viden.

Včasih se znajdejo pred nerešljivim jezikovnim problemom, ker jim slovar ne nudi dovolj izčrpne pomenske razlage in dovolj primerov rabe, zato brskajo po avtentičnih besedilih ali pa se obrnejo na lektorja, prevajalca. Njihova pričakovanja v zvezi s slovarjem slovenskega jezika so zato predvsem, da bi dal dovolj informacij o besedi, zato da bi jo lahko suvereno uporabili v kreativnem procesu iskanja in oblikovanja oglaševalskih zamisli. Zavedajo se tudi vpliva globalizacije na oglaševanje in v slovenščini vidijo tudi možnost odmika od poenotenih vzorcev, slovenski jezik s svojo kulturno specifiko jim daje možnost drugačnega, kreativnega razmišljanja. Hkrati pa opozarjajo na preglednost in ekonomičnost. Vse to povzemata tudi spodnji izjavi:

> Da je pregleden, funkcionalno zdizajniran, logičen za uporabo – 'simple and logic'.
>
> (informantka oblikovalka s statusom svobodne umetnice iz Kopra, rojena l. 1964)

> Morala bi se začutiti moč, teža posamezne besede. /…/ Več poudarka bi bilo potrebno dati kulturni specifiki, tudi v etimološkem pogledu, zato da bi oglaševalci več pomena posvečali ustreznemu prenosu globaliziranih oglaševalskih strategij in vsebin v slovensko kulturo.
>
> (informantka upokojenka, prej zaposlena v PR-službi večjega podjetja, iz Ljubljane, rojena l. 1946)

8 SKLEPNE UGOTOVITVE

Vse opazovane govorce in pisce, ki se z jezikom ustvarjalno ukvarjajo, nedvomno druži razvita in aktivna jezikovna zavest. Čeprav se večinoma ne ukvarjajo z metajezikovnimi vprašanji in tudi ne razmišljajo neposredno o tem, kakšni naj bodo jezikovni priročniki, pa pogosto razmišljajo o jeziku in ustreznem izražanju glede na okoliščine, v katerih se znajdejo, ter iščejo odgovore na različnih mestih, tudi v jezikovnih priročnikih.

Nedvomno pa je pri vseh opazovanih profilih, tj. literarnih ustvarjalcih, znanstvenikih, novinarjih in oglaševalcih, z vidika uporabe jezikovnih priročnikov opazna generacijska razlika. Predstavniki starejše generacije uporabljajo izključno tiskane jezikovne priročnike, srednja generacija se poslužuje tako tiskanih kot spletnih priročnikov, mlajša generacija pa uporablja izključno spletne jezikovne priročnike.

Vse skupine večinoma poznajo in uporabljajo precej klasične tiskane in spletne jezikovne priročnike (SP, SSKJ, slovarje tujk, dvo- in enojezične slovarje), le redki pa poznajo sodobnejše spletne portale z različnimi jezikovnimi viri in priročniki, kot so: Sporazumevanje v slovenskem jeziku, Portal jezikovnih virov, Termania, Fran, FB-portal Jezikovna Slovenija itd., korpuse pozna le ena intervjuvanka (gl. v Tabelah 1 in 2).

Tabela 1: Uporaba jezikovnih priročnikov.

Ali pri svojem delu uporabljate jezikovne priročnike?	**Literarni ustvarjalci**	**Znanstveniki**	**Novinarji**	**Oglaševalci**	Skupno
Da	6	10	5	5	26
Izjemoma	3	0	0	0	3
Ne	1	0	0	0	1
Skupno	10	10	5	5	30

Tabela 2: Uporaba spletnih jezikovnih priročnikov in orodij.

Ali poznate spletne jezikovne priročnike oziroma orodja?	**Literarni ustvarjalci**	**Znanstveniki**	**Novinarji**	**Oglaševalci**	Skupno
Da, poznam vključno s korpusi in jezikovnimi portali	0	2	1	1	4
Da, poznam različna orodja, a ne poznam korpusov in jezikovnih portalov	3	7	1	3	14
Da, a poznam samo spletne različice knjižnih izdaj	2	1	2	1	6
Nekatere poznam, a zelo redko uporabljam	3	0	0	0	3
Ne poznam	2	0	1	0	3
Skupno	10	10	5	5	30

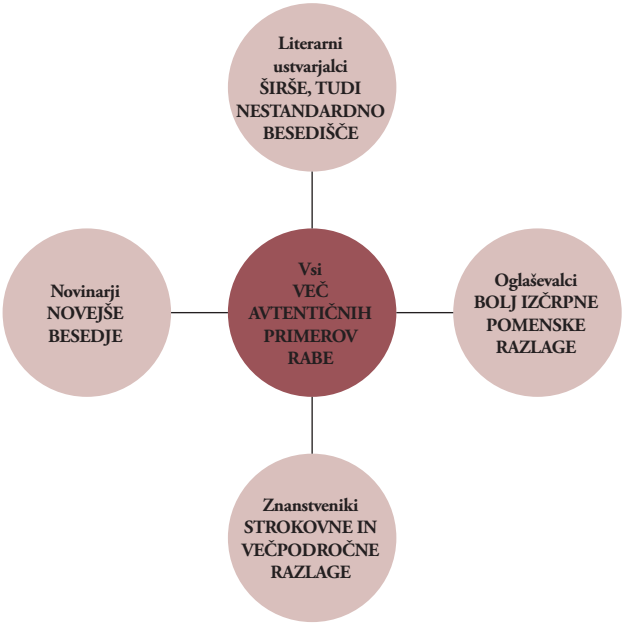
Jezikovne priročnike najpogosteje uporabljajo za iskanje pravopisnih in slovničnih informacij, pogosta sta tudi ugotavljanje natančnejšega pomena in iskanje primerov rabe, sledi zanimanje za novejše besedje, prevodne ustreznice v slovenščini, sopomenke in strokovne izraze (gl. Tabelo 3).

Tabela 3: Nameni uporabe jezikovnih priročnikov.

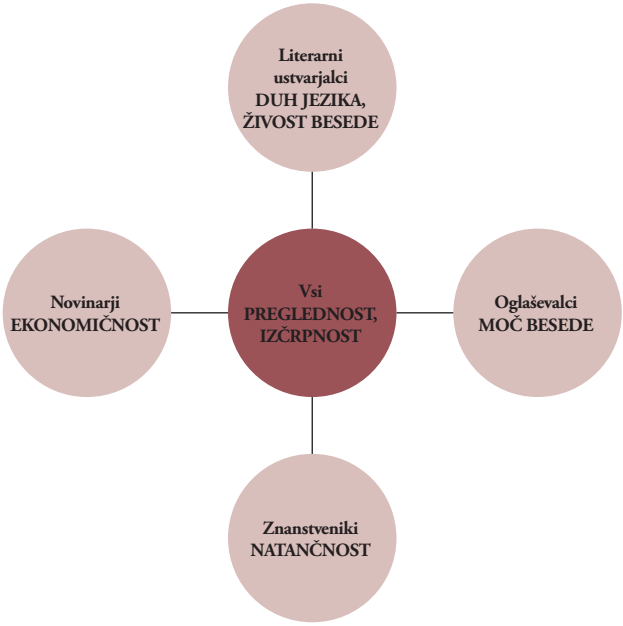
S kakšnim namenom oziroma ob katerih vprašanjih posežete po njih?	**Literarni ustvarjalci**	**Znanstveniki**	**Novinarji**	**Oglaševalci**	Skupno
Razlaga pomena	4	3	2	2	11
Natančnejši pomenski opis	6	8	0	5	**19**
Stilna zaznamovanost	3	0	2	4	9
Neologizmi, novejše besedje	2	6	5	3	**16**
Slovnična in pravopisna pravila	6	8	5	5	**24**
Sopomenke	4	6	0	3	**13**
Večpomenskost	1	2	0	0	3
Primeri rabe	5	6	3	5	**19**
Večpodročnost	0	5	1	0	6
Strokovni izrazi	0	9	3	0	**12**

S kakšnim namenom oziroma ob katerih vprašanjih posežete po njih?	**Literarni ustvarjalci**	**Znanstveniki**	**Novinarji**	**Oglaševalci**	**Skupno**
Prevodne ustreznice v slovenščini	0	10	2	2	**14**
Etimologija	1	2	0	2	5
Besedotvorne značilnosti	2	3	0	0	5
Naglas	2	0	3	0	5
Skupno	36/10=**3,6**	68/10=**6,8**	26/5=**5,2**	31/5=**6,2**	161/30=**5,3**

Glede pričakovanj, ki jih intervjuvanci imajo kot slovarski uporabniki, lahko pri vseh skupinah opazimo, da v obstoječih slovarjih – kot tudi v drugih jezikovnih priročnikih – občasno ne najdejo vseh odgovorov na svoja jezikovna vprašanja, predvsem pogrešajo izčrpnejše in avtentične primere rabe ter širši obseg zajetega besedišča, pogosto pogrešajo novejše in tudi nestandardno besedišče. Poleg tega vse skupine kot prvo lastnost, ki jo pričakujejo od slovarja, izpostavljajo preglednost. Najbolj izpostavljene elemente, ki jih intervjuvanci oziroma posamezne intervjuvane skupine ustvarjalcev pogrešajo v slovarjih in ki jih od njih pričakujejo, smo prikazali na Slikah 1 in 2.



Slika 1: Kaj jezikovni ustvarjalci pogrešajo v slovarju?



Slika 2: Kaj jezikovni ustvarjalci pričakujejo od slovarja?

Tako glede uporabe jezikovnih priročnikov kot glede pričakovanj slovarskih uporabnikov so, čeprav gre za majhen vzorec, zanimive nekatere hitro opazne razlike med skupinami. Iz Tabele 3, ki kaže število pojavitev posameznega namena uporabe jezikovnih priročnikov, lahko razberemo, da je za vsakega intervjuvanca v povprečju značilnih 5,3 pojavitev različnih namenov uporabe. Pri tem se kaže nadpovprečna raba slovarjev pri skupinah znanstvenikov (6,8 pojavitev) in oglaševalcev (6,2 pojavitev), blizu povprečja je skupina novinarjev (5,2 pojavitev), najmanj potreb po jezikovnih priročnikih pa so izkazali literarni ustvarjalci (3,6).

Prav tako lahko na osnovi celotne analize in slikovnega gradiva ugotovimo še podrobnejše značilnosti posameznih skupin. Tako kaže, da največ novejših spletnih jezikovnih portalov poznajo znanstveniki, ki jezikovne priročnike uporabljajo predvsem zaradi tvorbe in razumevanja strokovnega izrazja, zato si želijo čim bolj *natančnih razlag*, tudi z vidika različnih strok/področij. Raziskovalci tudi najbolj opozarjajo na interaktivnost in druge možnosti, ki jih za jezikovne priročnike nudijo nove tehnologije.

Pri uporabi le-teh za reševanje jezikovnih vprašanj jim sledijo oglaševalci, ki v jeziku in posledično v jezikovnih priročnikih iščejo predvsem spodbudo za svoje ustvarjalne zamisli, zato si želijo »čutiti« besede, njihovo *moč* in specifiko, tudi zato da z jezikom dosežejo odstopanje od prevladujočih globalizacijsko poenotenih oglaševalskih vzorcev.

Literarni ustvarjalci se delijo v dve skupini: so ali redni uporabniki jezikovnih priročnikov ali pa se zanesejo na svoj jezikovni čut in jezikovnih priročnikov skoraj ne uporabljajo; slednji so tudi edini, ki so priznali, da jezikovne priročnike zelo redko ali nikoli ne uporabljajo. Kljub temu pa se vsi živo zanimajo za jezik, začutiti želijo *duh jezika in živost besede*, česar si tudi želijo v slovarju slovenskega jezika.

Novinarji, ki občasno jezikovne priročnike sicer vsi uporabljajo, pa si zanje zaradi narave svojega dela ne morejo vzeti veliko časa, na kar opozarjajo tudi pri pričakovanjih glede slovarja slovenskega jezika – da bo pregleden in *ekonomičen*. Novinarji so tudi tisti, ki v slovarju najbolj pogrešajo novejše besedišče. Tako kot nekateri literarni ustvarjalci in oglaševalci se tudi nekateri novinarji zanesejo na lektorje, kar je razumljivo, glede na to, da imajo večji mediji, kot tudi založbe, organizirane lektorske službe.

Ob zaključku lahko rečemo, da opazovane profile jezikovnih uporabnikov in uporabnikov jezikovnih priročnikov lahko dejansko razumemo kot jezikovne ustvarjalce, ki jeziku in njegovim možnostim posvečajo veliko pozornosti. Prav zato bi morala temu njihovemu interesu slediti tudi jezikovna infrastruktura, ki bi lahko njihovo jezikovno zavest še razvijala in širila. Prvi pogoj za to pa je poznavanje njihovih interesov in potreb, zato bi moralo biti kontinuirano spremljanje uporabniških potreb ena glavnih nalog slovenskega slovaropisja.

S pomočjo uporabniških jezikovnih vprašanj in mnenj do boljšega slovarja

Špela Arhar Holdt, Jaka Čibej in Ana Zwitter Vitez

Abstract
In this paper, we examine the value of complementing existing dictionary user research by identifying and categorising authentic user-reported language problems. We present an analysis of 882 language problems and comments on the plans for a new monolingual dictionary of Slovene. The questions and comments were collected from four different language advice sites (two professional sites and two Facebook groups) and news article comment sections, and categorised manually. The results provide insights into how the dictionary should be designed in terms of content and interface in order to be as user-friendly as possible. The analysis of language questions has pointed to a need to provide dictionary users with clear guidelines on the choice and use of variants in a language, an interface that intuitively links different parts of the dictionary, additional information, e.g. on relevant language rules, and a manner of including the users in the process of dictionary creation. The analysis of user comments, on the other hand, has shown that in future lexicographic projects, it would be prudent to invest more resources in the communication with the interested public, to base lexicographic debates and decisions on the empirical data about (potential) dictionary users, and to present the use of public funds in a more transparent manner.

Key words: monolingual dictionary, user research, language problems, language questions, user comments

Ključne besede: enojezični slovar, slovarski uporabnik, jezikovne težave, jezikovna vprašanja, uporabniški komentarji

1 UVOD

Zmožnosti, potrebe in mnenja slovarskih uporabnikov[1] je mogoče raziskovati na različne načine. V metodološkem smislu se področje uporabniških raziskav poslužuje predvsem: (I) anket in intervjujev, ki tipično poizvedujejo, kako pogosto vprašani slovar(je) uporabljajo, za katere namene, kako v splošnem ocenjujejo uporabnost obstoječih priročnikov in kakšnih sprememb si želijo; (II) opazovanja slovarske rabe, kjer se z različnimi pristopi ugotavlja, katere informacije slovarski uporabniki iščejo, kako iskanje poteka, ali so bili podatki uspešno pridobljeni ali ne; (III) eksperimentov in testov, npr. za preverjanje, kako dobro testiranci slovarske podatke razumejo; in (IV) evalvacij posameznih slovarjev, npr. intuitivnosti in preglednosti uporabniškega vmesnika in podobno (več o tem v Arhar Holdt 2015).

Naštete metode lahko razkrijejo, katere priročnike slovarski uporabniki poznajo in uporabljajo, kako samoocenjujejo svoje potrebe, navade in želje glede (rabe) slovarja, kaj konkretno jim je pri določenem slovarju všeč in kaj ne, pa tudi, kaj z določenim slovarjem počnejo v procesu uporabe. Vse te informacije so nepogrešljiv temelj za razvoj slovarskih priročnikov in sodobni slovaropisni projekti jih uporabljajo v vedno večji meri. Čeprav so obstoječi pristopi k uporabnikom dragoceni za področje, pa so pomanjkljivi v smislu, da se sodelujoči (lahko) opredeljujejo le do obstoječega, poznanega stanja (Müller-Spitzer 2014: 169), s čimer se obstoječe stanje v večji meri ohranja kot spreminja. Na ta problem je opozoril W. Mentrup že leta 1984 in predlagal, da bi bilo slovaropisne uporabniške raziskave smiselno začeti korak pred slovarsko rabo, tj. pri jezikovnih motnjah oz. prekinitvah jezikovne rabe[2], ki do rabe slovarja (lahko) vodijo (Mentrup 1984, po Bergenholtz in Tarp 2004). Ideja je bila v tistem času v strokovnih krogih zavrnjena, danes pa je ponovno obujena v funkcijski teoriji slovaropisja (Fuertes-Olivera in Tarp 2014), ki slovar vidi kot orodje za reševanje uporabniških jezikovnih težav,[3] raziskave slednjih pa kot pot do (želenih) inovacij na ravni slovarske vsebine in oblike (Tarp 2009).

V pričujočem prispevku predstavljamo analizo uporabniških jezikovnih vprašanj in mnenj o slovarju, ki so bila objavljena v različnih spletnih okoljih (jezikovnih svetovalnicah, na omrežju Facebook in novičarskih forumih). S tem sledimo zgoraj predstavljeni ideji, pri čemer vpeljujemo pristop, ki na področju slovaropisja

1 V prispevku razlikujemo med pojmoma uporabnik in slovarski uporabnik. Prvo poimenovanje je splošnejše in se nanaša na uporabnike jezika oz. na uporabnike posameznih spletnih virov, kot je razloženo v nadaljevanju. Te uporabnike vedno razumemo kot potencialne slovarske uporabnike, vendar pa je o dejanskih slovarskih uporabnikih mogoče govoriti le ob obstoju podatkov, ki potrjujejo, da posameznik določen slovar oz. slovarje v praksi uporablja.

2 »/L/anguage-related disruptions in language use situtations« (prevod po Bergenholtz in Tarp 2004: 28).

3 O uporabni vlogi slovarja v slovenski družbi v primerjavi s simbolno in gradivno vlogo razpravlja M. Stabej (2009).

v raziskovalnem smislu do sedaj še ni bil preizkušen.[4] Razpravi o metodoloških vprašanjih, ki presega meje tega prispevka, se bomo posvetili na drugih mestih, v tej monografiji pa predstavljamo izsledke raziskave, predvsem tiste, ki lahko pripomorejo k prioritizaciji in organizaciji vsebin v (enojezičnem razlagalnem) slovarju ter k zasnovi in diseminaciji slovaropisnih projektov v slovenskem prostoru.

2 METODOLOGIJA

Uporabniška jezikovna vprašanja in mnenja je mogoče prestreči v primerih, ko so formulirana v jezikovnih svetovalnicah, na relevantnih forumih, skupinah na družbenih omrežjih in podobno. Zbiranje podatkov lajša dejstvo, da je gradiva veliko, saj je v slovenskem prostoru javna razprava o jezikovnih vprašanjih široko uveljavljena praksa (Kalin Golob 1996), s prehodom debate v spletno okolje (Žaucer in Marušič 2009) pa je gradivo postalo enostavneje dostopno in je bilo v slovenskem prostoru tudi že uporabljeno za raziskovalne namene. Hribar (2009) denimo analizira del jezikovnih vprašanj s foruma med.over.net v kontekstu razprave o uporabnosti obstoječih jezikovnih priročnikov.[5] Uporabo jezikovnih vprašanj, avtomatsko pridobljenih iz različnih spletnih okolij, za potrebe uporabnega jezikoslovja predstavlja priprava ontologije uporabniških normativnih zadreg (Bizjak Končar et al. 2011; H. Dobrovoljc in Krek 2011), ki je potekala pri projektu Sporazumevanje v slovenskem jeziku.[6] Projekt je razvil tudi dva predloga, kako identificirane jezikovne težave reševati: portal Slogovni priročnik (Krek 2012) in Pedagoški slovnični portal (Arhar Holdt et al. 2013). Na področju slovaropisja uporaba jezikovnih vprašanj in mnenj kot komplement obstoječim raziskovalnim metodam, kot rečeno, še ni bila preizkušena.

Za raziskavo smo zbrali vprašanja in mnenja iz različnih spletnih virov,[7] pri čemer so bili vključeni (I) viri, kjer na jezikovna vprašanja odgovarja strokovnjak, in viri, kjer so vprašanja zastavljena širši javnosti; (II) viri, pri katerih komunikacija poteka pisno, in viri, kjer je komunikacija govorna; (III) viri, kjer uporabniki zastavljajo jezikovna vprašanja, in viri, kjer izražajo svoje mnenje glede slovarja oz. slovaropisja.[8] Zaradi časovne potratnosti ročne kategorizacije podatkov smo

4 Čeprav lahko predvidevamo, da založniki spremljajo obiske svetovalnic, blogov in podobne aktivnosti svojih uporabnikov, tovrstni rezultati tipično niso objavljeni kot raziskave s področja slovaropisnih uporabniških raziskav, ampak uporabljeni interno za izboljšavo določenih izdelkov.

5 Analiza je pokazala, da jezikovna vprašanja odražajo številne pomanjkljivosti obstoječih jezikovnih priročnikov (nerazumljivost, nedoslednost, nedostopnost), pa tudi težave uporabnikov, da informacije ustrezno interpretirajo (Hribar 2009: 175).

6 Spletna stran projekta s povezavami na projektne rezultate: www.slovenscina.eu (dostop 8. 8. 2015).

7 Pri radijski oddaji Jezikovni svetovalni servis ne gre za izvorno spletni vir, vendar jo obravnavamo kot tako, ker smo do nje dostopali prek spletnega arhiva RTV Slovenija, http://4d.rtvslo.si/arhiv/ (dostop 8. 8. 2015).

8 V prispevku puščamo ob strani specifike spletnih virov, kjer so bila vprašanja objavljena, prav tako pa tudi odgovore, ki so jih na jezikovna vprašanja uporabniki dobili. Več o teh temah je mogoče prebrati v Kravos (2014), ki v primerjavo delovanja različnih jezikovnih svetovalnic vključuje tudi ŠUSS in radijsko oddajo Jezikovni svetovalni servis.

v prvem koraku raziskave lahko zajeli le omejen nabor virov; pri selekciji smo upoštevali tri kriterije: (I) enostavna dostopnost do podatkov (gl. predstavitev zbiranja podatkov spodaj), (II) heterogenost glede značilnosti, predstavljenih v Tabeli 1, in (III) stopnja ne/raziskanosti (preferirali smo do sedaj manj raziskane vire). Izven dometa raziskave so trenutno ostali nekateri pomembni viri, ki bi jih bilo nujno vključiti v naslednjem koraku, predvsem pogosto obiskana Jezikovna svetovalnica ISJFR ZRC SAZU in k jezikovnim vprašanjem usmerjeni forum Al' prav se piše … portala med.over.net.[9]

Tabela 1: Vključeni viri uporabniških vprašanj in mnenj.

Vir	Vrsta komunikacije	Medij	Vnos
Spletna jezikovna svetovalnica ŠUSS	Uporabnik zastavi vprašanje strokovnjaku, argumentirani odgovor sledi s časovnim zamikom.	pisni	jezikovno vprašanje
Radijska oddaja Jezikovni svetovalni servis	Uporabnik zastavi vprašanje strokovnjaku, odgovor dobi takoj in se lahko nanj tudi odzove.	govorni	jezikovno vprašanje
Specializirane skupine na omrežju Facebook	Uporabnik zastavi vprašanje skupini, ki združuje uporabnike glede na določen interes, posamezni člani skupine težavo komentirajo ter predlagajo rešitve.	pisni	jezikovno vprašanje
Novičarski forumi	Uporabnik deli svoje mnenje pod medijskim prispevkom na določeno vsebino, drugi bralci se lahko odzivajo, komentirajo.	pisni	mnenje, komentar

Pred nadaljevanjem je treba nekaj besed nameniti specifikam izbrane metodologije. Prvič, podatki, ki jih je mogoče dobiti s preučevanjem jezikovnih vprašanj, razkrivajo samo tiste uporabniške potrebe, ki so prepoznane (angl. *recognized needs* v Tarp 2009: 281) in glede katerih so se uporabniki odločili aktivno poiskati odgovor na določenem spletnem mestu; pri tem v raziskavi privzemamo, da uporabniki jezikovnih priročnikov pred poizvedbo niso uporabili oz. so jih, vendar neuspešno.[10] Podobne omejitve veljajo za mnenja, pridobljena z novičarskih forumov, ki jih zaradi specifičnosti ni mogoče posploševati na širšo populacijo. Drugič, pri uporabljenem postopku natančnejši podatki o uporabnikih niso pridobljivi,

9 Spletna stran Jezikovne svetovalnice ISJFR ZRC SAZU: http://isjfr.zrc-sazu.si/svetovalnica#v (dostop 8. 8. 2015) in foruma Al' prav se piše …: http://med.over.net/forum5/list.php?125 (dostop 8. 8. 2015).

10 Za jezikovna vprašanja, zastavljena na portalu med.over.net, N. Hribar predvideva, da »uporabnik sprašuje, ker odgovora ne pozna, in sicer: ker ga ne more najti v priročniku; ker priročnika nima pri roki; ker priročnikov ne uporablja ali pa po njih ne zna iskati; ker mu je laže postaviti vprašanje na internetu« (Hribar 2009: 175).

čeprav je o določenih značilnostih (npr. spol, okvirna starost, socialni status oz. poklic) pri nekaterih primerih mogoče sklepati iz uporabniške komunikacije ali splošnih značilnosti uporabljenega vira. Posledično je treba vzorec razumeti kot vprašanja in mnenja potencialnih in ne dejanskih slovarskih uporabnikov. Tretjič, v raziskavi niso bili zajeti vsi razpoložljivi viri vprašanj oz. mnenj, prav tako zajemanje ni potekalo vzorčeno oz. uravnoteženo po posameznih virih. Ta opozorila je treba upoštevati pri posploševanju rezultatov in pri primerjavah med potrebami in željami uporabnikov različnih virov (če bi te sploh imele smisel, saj ni nujno, da je posamezni uporabnik aktiven pri enem samem viru). Relativno obsežna količina podatkov kljub temu omogoča prve zaključke, ki bodo v nadaljnjih raziskavah preverjeni na večji količini gradiva.

Gradivo za raziskavo je bilo po večini zbrano ročno, razen (I) vprašanj jezikovne svetovalnice ŠUSS, ki so bila v sklopu projekta Sporazumevanje v slovenskem jeziku luščena avtomatsko (Bizjak Končar et al. 2011), za pričujočo raziskavo pa smo jih zgolj ponovno uporabili. (II) Gradivo iz radijske oddaje Jezikovni svetovalni servis je bilo zbrano tako, da smo izpisali vsa vprašanja iz devetih arhiviranih oddaj, ki so bile izvorno predvajane med letoma 2012 in 2013. (III) Vprašanja skupin z jezikoslovno tematiko na Facebooku (Društvo ljubiteljskih pravopisarjev in slovničarjev ter Za vsaj približno pravilno rabo slovenščine) so bila zbrana med novembrom 2014 in januarjem 2015. Iz vsake skupine smo zajemali gradivo med najnovejšimi objavami, dokler nismo dosegli 100 relevantnih objav (o tem več v nadaljevanju). (IV) Mnenja glede nastajanja novega slovarja slovenskega jezika smo zajeli na novičarskih portalih, ki omogočajo komentiranje uporabnikov (Rtvslo.si, Dnevnik.si, Delo.si in Družina.si). Zajeli smo komentarje pod objavami, ki obravnavajo Predlog za izdelavo Slovarja slovenskega jezika (2013), vlogo in nacionalni pomen slovarja, razpravo o sodelujočih institucijah in vprašanja o finančnih vidikih priprave slovarja.

Zbrana vprašanja in komentarji so bili ročno pregledani in kategorizirani glede na vsebino. Kategorije za označevanje so bile razvite na osnovi gradiva samega in se posledično pri označevanju vprašanj in komentarjev razlikujejo. Med kategorizacijo so bili iz nadaljnje obravnave odstranjeni za raziskavo nerelevantni vnosi, npr. objave, ki so vsebovale zgolj povezavo na članek ali spletno stran, komentarji, ki se niso nanašali na glavno temo pogovora (npr. zbadanje med posameznimi uporabniki), in objave, ki so bile vsebinsko preširoke (npr. *Ali je slovenščina najbolj zapleten jezik na svetu?*). Vsa ostala vprašanja smo ohranili, tudi če niso spraševala po informacijah, ki jih tipično prinašajo slovarski priročniki. Po kategorizaciji so podatki obsegali 882 vprašanj in mnenj, kot prikazuje Tabela 2.[11]

11 Postopek je bil uporabljen tudi za analizo vprašanj in odgovorov v Facebookovi skupini *Prevajalci, na pomoč!*, ki jo predstavlja prispevek Čibej et al. (2015).

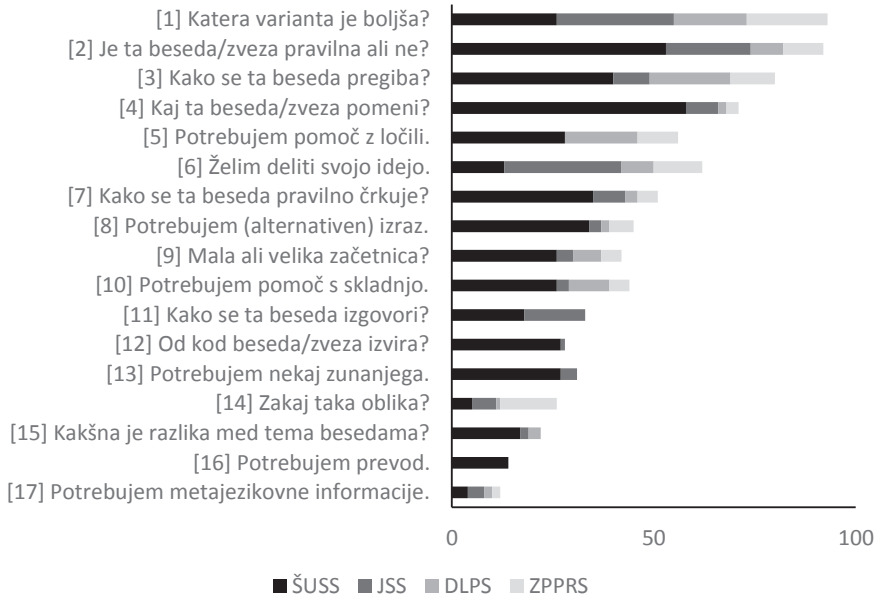
Tabela 2: Količina zajetega gradiva pred in po selekciji.

Vir	Izhodiščno št. enot	Št. kategoriziranih enot
Spletna jezikovna svetovalnica ŠUSS	508	451
Radijska oddaja Jezikovni svetovalni servis	143	143
Facebook: Društvo ljubiteljskih pravopisarjev in slovničarjev	114	103
Facebook: Za vsaj približno pravilno rabo slovenščine	119	103
Novičarski forumi	102	82
Skupno	**986**	**882**

V nadaljevanju predstavljamo rezultate kategorizacije, sledi pa jim strnitev nekaterih pomembnejših zaključkov, ki jih analizirani podatki nakazujejo.

3 UPORABNIŠKA JEZIKOVNA VPRAŠANJA

Graf 1 prikazuje rezultate kategorizacije jezikovnih vprašanj, ki smo jih zbrali s spletne svetovalnice ŠUSS, oddaje Jezikovni svetovalni servis (JSS) in skupin Društvo ljubiteljskih pravopisarjev in slovničarjev (DLPS) ter Za vsaj približno pravilno rabo slovenščine (ZPPRS). Podatki so v grafu prikazani združeno in z absolutno pogostnostjo.



Graf 1: Jezikovna vprašanja glede na pripisano kategorijo.

3.1 Ustreznost, pravilnost, raba

Kot je razvidno iz Grafa 1, je bila skupno gledano najpogostejša kategorija vprašanj *Katera varianta je boljša?*, pri kateri uporabnik izpostavi dve jezikovni različici, izvedeti pa želi, katera je v njegovem kontekstu ustreznejša (bodisi z vidika pravilnosti bodisi z vidika pomenske, kolokacijske ali slogovne ustreznosti).

[1] Potrebovala bi razlago besede mešati oz. povezave s tem: človek, ki meša neko stvar, je mešalec. Kaj je pa stroj, ki nekaj meša? Kontekst: imam pretočni mešalec ali pretočni mešalnik ali pretočno mešalo (stroj). [ŠUSS]

[2] Majhni stroški ali nizki stroški? [DLPS]

Podobna kategorija je *Kakšna je razlika med tema besedama?*, pri kateri se uporabnik prav tako odloča med dvema različicama, a ne sprašuje, katera je bolj ustrezna, temveč po razliki (npr. v pomenu, v rabi) med njima.

[3] Mi lahko kdo razloži razliko pri uporabi besed limonin in limonov? S kakšnim primerom, ki bo dvome razblinil za vse večne čase. [DLPS]

[4] Pojma HUMANITARNOST, HUMANOST: je pomen enega pojma popolnoma enak drugemu, kaj je pravilno, morebitne razlike, semantika; Konkreten pojem: začasno bivanje iz humanitarnih razlogov. [ŠUSS]

Druga najpogostejša kategorija je bila *Je ta beseda/zveza pravilna ali ne*? Ta kategorija je podobna kategoriji *Katera varianta je boljša?*, a se od nje razlikuje po tem, da uporabnika ne zanimajo informacije o pomenu ali ustreznosti, temveč zgolj potrditev, ali je nekaj v jeziku »dovoljeno« ali ne.

[5] Upam, da boste zlahka odgovorili na naslednje vprašanje: Kako je mogoče, da se tako množično (posebej moteče je to v medijih) uporabljajo izrazi 'zadane', 'prizadane' in podobno? To obliko rabe besede 'zadeti' sem prvič srečal nekje leta ’78, potem nekaj časa nič, zdaj pa praktično vsak dan. Saj ni pravilna, ne? [ŠUSS]

[6] Sken, skeniranje, skenirati ..... Je uporaba dopustna? Stari pravopis pravi, da ne. Kaj pa novi? [ZPPRS]

Pri kategoriji *Potrebujem (alternativen) izraz* gre tipično za uporabnika, ki išče v slovenščini nevtralnejši oz. ustreznejši ali »pravilnejši« izraz, npr. za nadomestilo določene narečne ali žargonske besede.

[7] Uf, ali ima »heksenšus« kakšno bolj pravilno različico v slovenščini? Povsem brez ideje. [DLPS]

[8] Ali obstaja kakšna slovenska beseda, ki bi ustrezala izrazu 'kasiranje'? ('Če potrebujete originalni račun, vas prosimo, da to poveste blagajničarki pred kasiranjem blaga ter predložite potrdilo o registraciji.') [ŠUSS]

3.2 Črkovanje in pregibanje

Sorodna vprašanjem o ustreznosti je kategorija *Kako se ta beseda pravilno črkuje?*, pri kateri uporabnik sprašuje po pravilnem zapisu besede.

[9] Kaj naredite z besedami, ki nimajo prevoda v slovenščino? Recimo milk shake - napišete fonetično milk šejk ali pustite v angleščini ? Ali kaj drugega? [DLPS]

[10] Zanima me, kako se pravilno napiše besedo japij. In definicija te besede! [ŠUSS]

V vseh štirih virih je bilo zelo pogosto povpraševanje po pregibnih vzorcih besed (kategorija *Kako se ta beseda pregiba?*). Velikokrat gre v takšnih primerih za pregibanje lastnih imen (zlasti tujih):

[11] A ma kdo kako idejo, kako se sklanja »dao« (to je temeljni pojem daoizma oz. taoizma)? [ŠUSS]

[12] Ob poročanju o dogajanju v državi Burkina Faso sem zasledila dva načina sklanjanja: v Burkini Faso in v Burkina Fasu? Drugi način mi zveni narobe, ampak maybe it's just me. Res pa je, da je pridevnik burkinafaški... [ZPPRS]

Pri povpraševanju po pregibnih vzorcih uporabniki pogosto želijo tudi razlago oziroma splošno pravilo, ki ta pregibni vzorec določa (zlasti pri specifičnih lastnih imenih). Na ta način želijo učinkovito rešiti nekaj, kar v jeziku dojemajo kot nelogično, nedosledno ali nejasno. Enako je tudi pri sicer manj pogosti kategoriji *Zakaj taka oblika?*, pri kateri uporabnik ustrezno obliko že pozna, ne razume pa, zakaj je takšna, kot je, zato prosi za pojasnilo.

[13] Vanilja. (Pogojno vanilija z dodatnim i). Vsekakor pa ženski spol. Zakaj za vraga potem v kuharskih knjigah piše vanilijev sladkor, namesto vaniljin sladkor (ali vanilijin)? Je to enako kot spodaj lipov čaj/lipin čaj, breskov sok, bananov sok? [ZPPRS]

[14] Zakaj gremo na žago in ne k žagi? [JSS]

3.3 Pomen in izvor

Vprašanja v zvezi s pomenom in izvorom besede ali besedne zveze so bila najpogostejša pri ŠUSS-u, pri ostalih virih pa manj.

[15] Zanima me kaj pomeni »vertikalni« in kaj »horizontalni« koncern? Na to sem naletela pri svojem projektnem delu o Intereuropi. In sicer pri združevanju Intereruope z Luko Koper. Združili naj bi se v VERTIKALNI KONCERN! [ŠUSS]

[16] Zanima me, če veste kaj o izvoru besede »kodrlajsast« (pač v pomenu pisan, raznobarven, šarast), in o tem, v katerih narečjih je razširjena. [ŠUSS]

3.4 Izgovor in ločila

Kategoriji *Kako se ta beseda izgovori?* in *Potrebujem pomoč z ločili* sta se izkazali za nekoliko bolj specifični od ostalih, saj sta vezani na govorni in pisni medij. Vprašanj o izgovoru je bilo tako največ pri radijski oddaji Jezikovni svetovalni servis, o ločilih pa so uporabniki povpraševali v ostalih (pisnih) virih. Pri vprašanjih o ločilih so najpogostejše vejice, uporabniki pa tudi v tem primeru pogosto prosijo za razlago v obliki pravila.

[17] Izgovorjava deL, namesto deU. Nas so učili v OŠ, da se L uporablja na koncu besede samo pri tujkah. [JSS]

[18] Mi lahko prosim nekdo razloži, zakaj se v primeru: »Pozimi in kadar je mraz, moramo zakuriti.« piše vejica. [DLPS]

3.5 Izražanje mnenja

Za pogosto se je izkazala tudi kategorija *Želim deliti svojo idejo*, v katero smo razvrstili vprašanja, pri katerih želi uporabnik izraziti svoje mnenje o jezikovni rabi. Tovrstna vprašanja se lahko osredotočajo na zelo specifičen jezikovni element ali pa na splošno jezikovno rabo.

[19] Radijci na prvem programu – začela se je uveljavljati angleška beseda vesel ali dober vikend. Dajte to prenehat, ne mešajte angleških popačenk v ta besednjak. Sicer bo ta ful kul generacija v prihodnosti pela Zdravljico v angleščini. Celo doktorji znanosti govorijo in rečejo deset besed, tri so tujke, tri so angleške, tri popačenke, sej komi še kej ostane. Slovenščina je ogrožena, res, na žalost. [JSS]

[20] Če prevajamo v slovenščino imena vladarjev, papežev in svetnikov, bi bilo po mojem logično in lepo, da bi tudi imena članov kraljevih družin, npr. princ Karel in princ Andrej. Tudi španski kralj nam je očitno kulturno veliko dlje kot angleška kraljica, sicer bi bil Ivan Karel in ne

> Juan Carlos. Seveda kakšne kraljice Beatrix ne gre prevajati, saj tudi s tankom ne pridemo na Triglav. Naprej se še čudim, da za židovska in zlasti arabska imena še vedno nimamo enotne transkripcije, čeprav se dosti pojavljajo v medijih. Včasih Mohamed, včasih Muhammad. [ŠUSS]

Ta kategorija je bila najpogostejša pri vprašanjih iz oddaje Jezikovni svetovalni servis, v manjši meri pa se je pojavljala tudi pri drugih virih. Angažiranost uporabnikov priča o tem, da čutijo potrebo, da izrazijo svoje mnenje in poskušajo na tak način vplivati na jezikovno rabo, zlasti v primerih, ko jo dojemajo kot napačno ali motečo.

3.6 Metajezikovne in druge dodatne informacije

Čeprav redkeje, uporabniki sprašujejo tudi po metajezikovnih informacijah, predvsem v svetovalnici ŠUSS pa tudi po drugih informacijah, npr. raznih zanimivostih in statistikah o slovenščini ali dostopnosti jezikovnih virov.

[21] Zanima me, pod katero besedno vrsto uvrstiti besedico ne (ne znam). Po SS besedica ne ne sodi pod glagol (SS, 449), v nekaterih vadnicah (M. P. Povodnik, Oblikoslovje) pa besedo ne najdemo pri glagolu. Popolna zmeda! [ŠUSS]

[22] Enkrat ste mi že pomagali iz zagate, zdaj me pa zanima, če mi lahko poveste, katera slovenska beseda je najdaljša. Sem namreč varuška (au pair) v Franciji, kjer skrbim za dve nadobudni šolarki (1. in 3. razred), ki že znata šteti po slovensko in nekaj drugih besed, zdaj ju pa zanima najdaljša slovenska beseda. [ŠUSS]

3.7 Ostale kategorije

Omeniti je treba še kategoriji *Potrebujem pomoč s skladnjo* in *Mala ali velika začetnica?*. Za pripravo slovarja se na prvi pogled zdita manj relevantni, vendar omogočata razmislek o povezovanju slovarskih informacij s skladenjsko in normativno ravnijo.

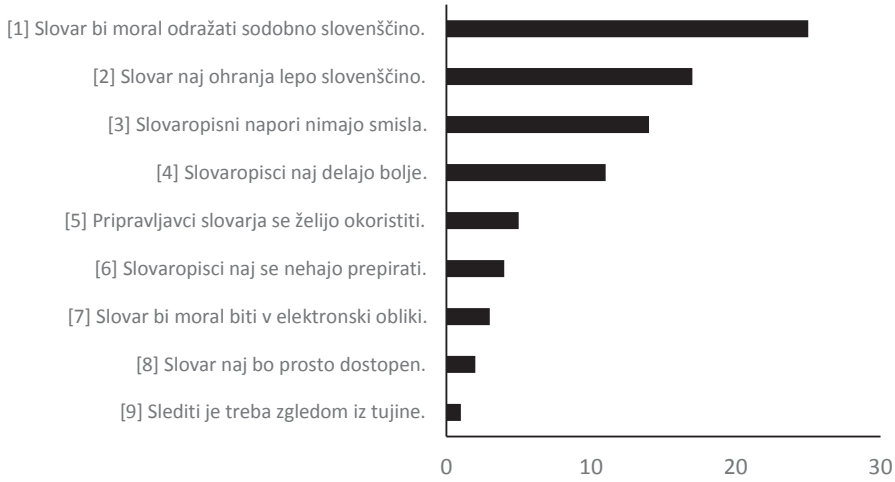
[23] Dolg v višini SLABE 4 mio EUR ali - Dolg v višini SLABIH 4 mio EUR? [DLPS]

[24] Pišemo g./ga./gdč. pri naslovu na pismu z veliko ali z malo začetnico? Na primer: G./g. Janez Novak, Slovenska 1, Maribor [ZPPRS]

4 MNENJA UPORABNIKOV O SLOVAROPISNIH PROJEKTIH

Graf 2 prikazuje rezultate kategorizacije uporabniških komentarjev na novičarskih portalih, ki so objavljali prispevke s slovaropisno tematiko in omogočajo komentiranje uporabnikov (Rtvslo.si, Dnevnik.si, Delo.si in Družina.si). Kot je razvidno iz grafa, smo komentarje uporabnikov razvrstili v devet kategorij, ki slovaropisno razpravo osvetljujejo s štirih krovnih vidikov:

- prioritete (kategoriji 1 in 2 izpostavita vprašanje, ali bi moral slovar odsevati predvsem sodobno ali standardno/knjižno/pravilno slovenščino),
- institucije (kategorije 4, 5, 6 in 9 razkrivajo odnos uporabnikov do slovaropisnega dela in sodelujočih institucij),
- tehnični vidiki (iz kategorij 7 in 8 je razvidno mnenje uporabnikov o mediju in dostopnosti slovarja ter finančne vidike njegove priprave),
- smiselnost razprave (kategorija 3 osvetljuje za uporabnike pomembnejše teme).



Graf 2: Mnenja in komentarji glede na pripisano kategorijo.

4.1 Prioritete

Prioritete uporabnikov se kažejo skozi dve najmočneje zastopani kategoriji: pri prvi kategoriji uporabniki zavzemajo stališče, da bi moral slovar odsevati sodobno slovenščino (30 % vseh komentarjev), pri drugi kategoriji pa uporabniki trdijo, da bi moral slovar predstavljati sredstvo za ohranjanje standardne (oz.

knjižne, »lepe«) slovenščine (21 % vseh komentarjev). Ker gre za tradicionalno nasprotujoči si stališči, ne preseneča, da uporabniki v svojih komentarjih ne izražajo zgolj mnenja, ampak se čustveno odzivajo, zahtevajo, opozarjajo, se šalijo in ostro kritizirajo.

Uporabniki, ki se zavzemajo za vključevanje sodobne jezikovne rabe v slovar, z zaznamovanimi izrazi (*grammar-nazi*) označujejo nasprotno stališče:

[25] le neki »grammar nazi-ji« bi radi narod učili »pravilno«, kot da brez njih ne gre. A se razumemo med seboj? razumljivost je glavni predpogoj vsega

Uporabniki, ki zagovarjajo rabo standardne slovenščine v slovarju, izpostavljajo nestrinjanje z vključevanjem dialektalnih izrazov (*sma*), slenga (*kva*) in angleških besed (*flat*) v slovar, kritizirajo pa tudi jezikovno rabo v novih medijih in med priseljenci:

[26] Če ne bi bilo takšne »kulturne« izmenjave kot je bilo takoj po drugi svetovni vojni med jugoslovanskimi narodi, bi za moje pojme slovenščina dobila boljši položaj.

4.2 Sodelovanje institucij

S komentarji, usmerjenimi v sodelovanje različnih institucij pri preteklih in predlaganih slovaropisnih projektih (26 % vseh komentarjev), uporabniki ocenjujejo dosedanje delo strokovnjakov in opozarjajo na nevarnost, da bi se posamezniki lahko okoristili s proračunskimi sredstvi. Ta skupina komentarjev odseva negativno mnenje in nizko stopnjo zaupanja tako v institucije, ki tradicionalno izvajajo slovaropisno dejavnost, kot tudi v institucije, ki so se na novo vključile v razpravo o novem slovarju:

[27] Podpiram vzdrževanje slovenskega jezika, ampak 4,2 mio pa je preveč! Jasno je, da se tu pasejo na davkoplačevalskem denarju. Preletet star slovar pa kaj samo dodat ali spremenit vsekakor ni vredno toliko. Evo jas dam ponudbo, za 10 000€ pa bi se potrudil :)

Kritike so največkrat usmerjene k oblikovanju dveh nasprotujočih si konzorcijev, zaradi katerega se oddaljuje operativna priprava slovarja. Uporabniki se pogosto šalijo na račun nesoglasij med akterji:

[28] Če bodo preveč mleli in čakali bo naslovnica izgledala tako: Речник сл овеначког језика.

4.3 Tehnični vidiki

Tehnični vidiki slovarja zajemajo 7 % vseh komentarjev ter se nanašajo na medij in dostopnost slovarja. Večina uporabnikov je mnenja, da bo slovar optimalno izkoriščen v elektronski obliki:

[29] SSKJ na spletu = uporabljali ga bodo vsi. Od Goričkega do Pirana, od Kočevske reke do Triglava. SSKJ v tiskani različici = nabiral bo prah. /.../ .

Komentarji, povezani z dostopnostjo slovarja, so dobronamerni in konstruktivni, večinoma pa izražajo enotno mnenje, da je smiselna prihodnost slovarja pogojena s prostim spletnim dostopom:[12]

[30] in ja za najboljšo ohranitev/uporabo slovenskega jezika mora biti slovar ZASTONJ, ...v CD žgoščenki na vsak dom! Financirati mora država.

4.4 Smiselnost razprave

Zadnji vidik analize odzivov uporabnikov razkriva mnenje, da slovaropisni napori nimajo pravega smisla (16 %). Na podlagi vsebine komentarjev sklepamo, da gospodarske razmere v Sloveniji močno nižajo motiviranost uporabnikov, da bi sprevideli smiselnost finančnega podpiranja slovenskega slovaropisja:

[31] Sprašujem se ali bomo Slovenci čez pet let sploh še rabili slovar slovenskega jezika. Država je sesuta in že danes se premnogi Slovenci podajajo s trebuhom za kruhom, zato je vsakega evra, ta trenutek škoda, za razne znanstvene teorije naše »inteligence«.

Komentarji, povezani z nesmiselnostjo slovaropisne razprave, večinoma ne prinašajo konstruktivnih predlogov za izboljšanje koncepta novih slovarskih priročnikov, temveč predstavljajo pesimistične scenarije o izginjanju slovenščine zaradi gospodarskih in globalizacijskih razlogov:

[32] Ma...., slovenscini slabo kaze zato, ker me se blagajnicarka pri Mercatorju na vicu ne razume. Torej kaj mi bo jezik, ki ga se polovica Slovencev ne razume?

12 Večina analiziranih komentarjev je bila zajetih pred izidom slovarja SSKJ2.

5 OD UPORABNIŠKIH VPRAŠANJ IN MNENJ K SPREMEMBI SLOVARJA OZ. SLOVAROPISNIH PRAKS

V nadaljevanju predstavljamo nekaj glavnih ugotovitev raziskave, pri čemer smo za predstavitev izbrali samo tiste izsledke, ki nakazujejo potrebo po spremembi trenutnega razumevanja slovarja in slovaropisja v slovenskem prostoru.

5.1 Uporabniki pogosto želijo primerjavo dveh ali več jezikovnih možnosti.

V analiziranih jezikovnih vprašanjih se pogosto pojavlja želja po primerjavi dveh ali več konkurenčnih jezikovnih možnosti, npr. primerjava pomena oz. značilnosti rabe pomensko (*slika – fotografija*) ali oblikovno podobnih besed (*komuniciranje – komunikacija*), primerjava različic določenega termina (kurent – korant) oziroma pravilnosti/nevtralnosti/ustreznosti jezikovnih variant na fonetični (*áneks – anéks*), oblikoslovni (*zadane – zadene*), leksikalni (*smatrati – meniti*) ali skladenjski (*čez cesto – prek ceste*) ravni. Ena od prioritet sodobnega digitalnega slovarja bi torej morala biti zasnova, ki želene primerjave v čim večji meri omogoča. Prikaze sinonimije in zgledov rabe, ki jih omogoča tiskana različica slovarja, bi bilo v digitalni obliki denimo mogoče preseči z naprednim povezovanjem med slovarskimi enotami in možnostmi za njihovo primerjavo, npr. z vpeljavo vzporednega iskanja in pogleda v dve ali več slovarskih gesel hkrati.

5.2 Kadar obstajajo v jeziku variante, želi uporabnik jasno opredelitev razlik.

Na prejšnjo točko se veže tudi ugotovitev, da v primerih, ki se ukvarjajo z vprašanjem variantnosti, uporabniki želijo jasne in eksplicitne smernice, ki jim pri izbiri pomagajo, pri čemer jih praviloma zanima nevtralnejša jezikovna izbira (npr. kateri od danih terminov je najbolj uveljavljen, katera je standardna alternativa za določeno nestandardno besedo, katera od dveh oblikovnih variant je pravilnejša/ustreznejša/boljša). Na tovrstna vprašanja v tradicionalnem slovarskem opisu odgovarjajo kvalifikatorji, ki jih je v sodobnem digitalnem slovarju mogoče preoblikovati oz. dopolniti z vključitvijo in vizualizacijo različnih podatkov iz jezikovne rabe (frekvenca, žanrski podatki ipd.).

5.3 Uporabnikov ne zanima samo, kaj je v jeziku pravilno/ustrezno, ampak tudi, kaj ni.

Uporabniki pogosto želijo eksplicitno informacijo o tem, ali je določen jezikovni element, ki so ga v jezikovni rabi spoznali za problematičnega, skladen z obstoječo kodifikacijo ali ne, pri čemer se pojavljajo tudi napačne interpretacije slovarskih podatkov oz. neustrezno razumevanje slovarja kot jezikovnega priročnika, npr. uporabniške predpostavke, da če nečesa ni v slovarju, ni pravilno, oz. da je vse, kar je v slovarju, (enako) ustrezno, ne glede na podatke o značilnostih rabe. Predvidljivo je, da bo napačnega interpretiranja glede na (ne)obstoj v digitalnem slovarju, ki se hitro posodablja in spremlja uporabniške vsebinske komentarje, manj. Željam po opredelitvi nepravilnega/neustreznega pa je v določeni meri mogoče zadostiti z vizualizacijo standardnih in nestandardnih jezikovnih izbir oz. vključitvijo in opisom tipičnih jezikovnih napak.

5.4 Uporabnike zanima tako standardno kot nestandardno v jeziku, tako splošno kot specializirano besedišče in tako občna kot lastna imena.

Spraševanje o pravilnosti in nevtralnosti je pogost del jezikovnih vprašanj, v katerih uporabniki izbirajo med več jezikovnimi možnostmi. Ko sprašujejo o informacijah glede posamezne besede, pa jih zanima tako standardno kot tudi nestandardno besedišče. Slednje se sicer pojavlja bistveno redkeje, gre denimo za vprašanja o pomenu, izvoru ali razširjenosti določene narečne besede (*gautre, gmajna*), ali pa uporabniki iščejo standardno sopomenko (*heksnšus, felš*). Na drugi strani je mogoče zapisati, da uporabnike zanima tako splošno kot specializirano besedišče: vprašanj o terminologiji je precej, pojavljajo se na ravni pomena (*inundacija*), pregibanja (*kava kava*), črkovanja (*skanogram – skenogram*) ter, kot že rečeno, pri iskanju najboljše različice (*frezanje – rezkanje*) ali podomačene ustreznice (*kasiranje*). Zadnja skupina, ki jo velja omeniti, so še lastna imena, ki uporabnike zanimajo s stališča zapisa z veliko/malo začetnico, pregibanja, besedotvorja (svojilnih pridevnikov in imen prebivalcev), izgovora, pa tudi pomena in izvora besede.

5.5 Jezikovna vprašanja ne odražajo nujno slovarske delitve leksikalnih informacij.

Jezikovna vprašanja niso zgolj enoznačne poizvedbe po posamezni točno določeni jezikovni informaciji, ampak lahko sprašujejo o več primerih hkrati ali

po podatkih, ki so v slovarjih običajno predstavljeni v različnih razdelkih.[13] S tem se potrjuje opozorilo W. Mentrupa, da je napačno predpostavljati, da situacije slovarske rabe sistematično korelirajo s slovarskimi razdelki, saj so ti aposteriorni slovaropisni konstrukt (Mentrup 1984: 151 po Bergenholtz in Tarp 2004: 26–27). Slovarska podoba, ki želi biti za reševanje jezikovnih problemov intuitivna, se mora torej izogibati pretirani delitvi informacij v (poimenovane) razdelke in namesto tega izhodišče navigacije zasnovati iz preglednega izhodiščnega nabora informacij, ki jim uporabnik lahko sledi z nadaljnjimi kliki do podrobnosti.[14]

5.6 Ko gre za vprašanja o izgovoru, uporabniki po tonemskosti ne sprašujejo.

Za izbiro prioritet slovarskega projekta ni relevanten le podatek, po katerih informacijah uporabniki najpogosteje sprašujejo, ampak tudi, po katerih ne sprašujejo. V analiziranih podatkih to velja za tonemski naglas, o katerem ni vprašal niti en uporabnik. Odsotnost vprašanj o tonemskem naglasu nakazuje, da je ta jezikovna značilnost pri uporabnikih manj uzaveščena oz. redko (če sploh) povzroča motnje v jezikovni rabi.

5.7 Nekatere vrste jezikovnih vprašanj so tesneje vezana na prostočasno rabo slovarja.

Kadar uporabniki v jezikovnih vprašanjih razkrijejo situacijo, ki je k vprašanju vodila, je mogoče vprašanje umestiti v kontekst izobraževanja, poklicne rabe ali prostočasne rabe. Pri večini od kategorij, ki jih predstavlja Graf 1, je mogoče najti v tem smislu heterogene primere. Izstopajo pa vprašanja o izvoru besede ter različnih jezikovnih statistikah (najtežje izgovorljiva slovenska beseda, seznam najpogostejših besed), za katere se zdi, da so izraziteje vezana na prostočasno rabo.

13 Nekaj primerov iz svetovalnice ŠUSS: *Zanima me izvor in pomen besede oz. slovenski izraz za "gautre". // Sem študent in v krogu svojih prijateljev in sošolcev veliko uporabljamo besedo SKRIPTA. Zanima me, kako se ta beseda pravilno sklanja in v kakšnem smislu je raba te besede sploh upravičena. // Kako se sklanja Institut Pasteur (v Parizu)? Moja logika: ali rečeš 'na Institut Pasteur' ali pa 'na inštitutu (Institute) Pasteur'. Ugovora: (a) moj šef napiše na Institutu Pasteur (b) kaj pa sklanjanje 'na Institutu Jožef Stefan'? In druga stvar, ostaja mi ta problem s/š. Ne vem, kateremu je institut za bruhat, ampak IJS je pa še vedno Institut in ne Inštitut // Bolnica in bolnišnica (je bolnica, le ženska, ki je bolana in bolnišnica ustanova za zdravljenje ? Prometni znaki v Ljubljani nam kažejo pot v Bolnico, je torej to pravilno?). Je izraz bolnica, ki ga uporablja nevedno ljudstvo v pomenu ustanove pravilno? Ali je res, da se, če se že izgovarja BOLNICA, izgovori vsaj 'bounica', torej v izg. kot u? Torej ali nasploh drži, da ljudstvo ne pozna pravil za izgovorjavo, ali obstajajo (vem, da obstajajo toda, kje jih naj dobim napisane ?) torej pravila, kdaj se izgovarja v kot u in kdaj ne?*

14 V slovenskem prostoru je podobna ideja implementirana na portalu Termania, in sicer kot prikaz skrajšanih gesel v nizu iskalnih zadetkov, kjer so za reprezentativne izhodiščne informacije izbrane iztočnica in primeri prevodov oz. definicij (Romih in Krek 2012). Za enojezični razlagalni slovar bi bilo nabor izhodiščnih informacij mogoče zasnovati in oblikovati s pomočjo rezultatov pričujoče analize.

Ta podatek lahko pomaga pri prioretiziranju slovarskih vsebin in pri razvrščanju podatkov, če privzamemo, da je za nekatere informacije pomembno, da so uporabniku na voljo čimprej, druge pa so lahko na voljo na klik.

5.8 Uporabniki sprašujejo tudi po informacijah, ki niso tipično slovarske.

Digitalna oblika slovarja omogoča vključevanje novih vrst vsebin, npr. slikovnega gradiva, zvočnih datotek ali videoposnetkov, različnih motivacijskih vsebin za uporabnike, na drugi strani pa povezovanje slovarskih informacij navzven, npr. z besedilnimi korpusi, enciklopedijami, jezikovnimi svetovalnicami itd.[15] Analiza jezikovnih vprašanj je v zvezi z novimi možnostmi pokazala dve pomembni tendenci: (I) uporabniki sprašujejo po razlogih, pravilih, ki so podlaga določene jezikovne odločitve,[16] in (II) pomemben del komunikacije o jezikovnih dilemah je možnost izražanja lastnega mnenja, ideje oz. predloga za spremembo. Na osnovi podatkov bi bilo torej med prioritetami mogoče izpostaviti povezavo slovarskih informacij z informacijami o jezikovnih pravilih ter vključitev uporabnikov v slovaropisni proces, o čemer je več govora tudi na drugih mestih monografije.

5.9 Uporabniki želijo prosto dostopen spletni slovar in transparentne slovaropisne projekte.

Medij slovarja v mednarodnih slovaropisnih praksah že dolgo ni več vprašanje (Rundell 2014), analiza uporabniških komentarjev na novičarskih portalih pa je pokazala, da uporabniki izražajo enotno željo tudi po prosto dostopnem spletnem slovarju. Enako poenoteno so uporabniki pokazali nizko stopnjo zaupanja tako v institucije, ki tradicionalno izvajajo slovaropisno dejavnost (*brontozavrske institucije*), kot tudi institucije, ki so se pred kratkim vključile v razpravo o novem slovarju (*Podpiram vzdrževanje slovenskega jezika, ampak 4,2 mio pa je preveč*). Ti rezultati nakazujejo potrebo, da se ob pripravah slovaropisnih projektov več energije vloži v komunikacijo z zainteresirano uporabniško javnostjo, empirično utemeljevanje strokovnih odločitev in transparentno predstavljanje upravičenosti porabe sredstev.

15 Nekaj primerov slovaropisnih idej o vključevanju multimedije in povezovanju z izvenslovarskimi podatki predstavlja de Schryver (2003: 165–172).

16 Nekaj primerov iz svetovalnice ŠUSS: *Vhleviti ali uhleviti (tako je v SSKJ). Saj ne gremo u hlev temveč v hlev. Morda na u izgovarjamo, pisati pa bi se po mojem mnenju moralo na v.* // *Kljub temu, da sem prebral vaš odgovor o rabi velike začetnice v geografskih imenih (omenjenega odgovora na strani ne boste našli, op. ur.), se srečujem s problemom pri pisanju imena Ptujske Toplice (enako Dolenjske Toplice ...). Prosim, da mi razložite, zakaj je v tej zvezi TOPLICE zapisano z veliko, saj po mojem mnenju spada v isti koš kot mesto, vas ...* // *Iz daljnega Stuttgarta vam pošiljam naslednje vprašanje, ki me muči že nekaj mesecev. Gotovo ste že vsi slišali za* ***Feng Shui****. Če ne, pobrskajte po knjigah. Moje vprašanje: kako se sklanja samostalnik Feng Shui? s Feng Shuijem? s Feng Shujem? pri Feng Shuiju? ... Še ena opomba: beseda Feng Shui se po pravilu izgovori 'fangšvei' mogoče celo 'fang švej'. Vpliva izgovorjava tudi na pisavo pri sklanjatvah?*

5.10 Dihotomija *splošni* vs. *zahtevni uporabnik* zahteva nov razmislek.

V stroki se slovarske uporabnike tipično deli na »splošne« in »zahtevnejše«[17], ki pričakujejo določene dodatne značilnosti na ravni slovarske vsebine in oblike. Kot zahtevne uporabnike se običajno vidi jezikoslovce, lektorje, prevajalce in podobne uporabniške skupine. Vendar pa so Čibej et al. (2015) pokazali, da prevajalci svoje jezikovne dileme in pričakovanja v odnosu do jezikovnih priročnikov formulirajo realno, natančno in dokaj konsistentno, njihovi predlogi pa so relevantni tudi za druge uporabniške skupine. Analiza komentarjev na novičarskih portalih pa je pokazala nabor raznorodnih, nasprotujočih si in včasih precej nerealnih pričakovanj oz. zahtev (npr. finančna ocena uporabnika, ki pravi, da *za 10 000€ pa bi se potrudil*).[18] Zato je treba v prihodnje v strokovni debati natančneje opredeljevati uporabniške skupine in pred ločevanjem povprečnih in zahtevnih uporabnikov in posledično povprečnih in zahtevnih slovarskih lastnosti »naslovnike in uporabnike slovarjev o vsem tem tudi kaj vprašati« (Logar 2009: 230).

6 SKLEP

Prispevek predstavlja rezultate prve analize uporabniških jezikovnih vprašanj ter uporabniških komentarjev o aktualnem dogajanju na področju slovenskega slovaropisja. Predstavljeno delo prinaša podatkovno podstat za razprave o jezikovnih in slovarskih potrebah uporabnikov in uporabnic slovenščine in preizkuša postopek, kako bi bilo tovrstne podatke mogoče raziskovati. S tem dopolnjujemo metodologijo uporabniških raziskav za potrebe slovaropisja v smeri, ki jo napovedujejo premisleki o vlogi slovarja v družbi kot vlogi, ki mora biti usmerjena tudi in predvsem k reševanju uporabniških jezikovnih težav (Tarp 2009; Stabej 2009).

Kot je bilo izpostavljeno v razdelku Metodologija, je treba rezultate razumeti z določeno mero previdnosti. V nadaljevanju bi bilo smiselno analizirati podatke še iz drugi relevantnih virov in jih primerjati s prvimi rezultati, pri čemer bi bilo postopek na ravni pridobivanja in kategoriziranja vprašanj treba vsaj delno avtomatizirati. Tudi kategorizacija zahteva ponovni premislek, saj se je zaradi kompleksnosti vprašanj in mnenj kot večji problem postopka izkazala nezadostnost enoznačnega kategoriziranja. Predvideva se tudi, da se bo z vključitvijo novih podatkov pojavila potreba po širitvi obstoječega nabora kategorij.

17 V posvetu o novem slovarju (Perdih 2009) avtorji uporabljajo tudi izraze *laični*, *navadni* in *povprečni* uporabnik.

18 Čeprav ti uporabniki morda niso dejanski slovarski uporabniki, niti kot populacija splošno reprezentativni, najbrž ustrezajo trenutnim ohlapnim kriterijem ciljnega uporabnika enojezičnega razlagalnega slovarja (gl. razpravo Arhar Holdt 2015).

Vpogled v avtentične jezikovne zadrege nakazuje točke, v katerih je mogoče novi slovarski priročnik zasnovno, vsebinsko in oblikovno najbolje približati uporabniku in njegovim jezikovnim potrebam. Analize, kakršna je predstavljena v tem prispevku, je mogoče upoštevati na ravni prioritizacije in organizacije vsebin pri digitalnem enojezičnem razlagalnem slovarju za slovenščino oz. pri sami zasnovi in diseminaciji slovaropisnega projekta kot celote. Obračanje k uporabniku jezika in situacijam, v katerih uporabnik slovarja verjetno ni uporabil, lahko pa ga bi – ali pa ga je uporabil, vendar neuspešno – je toliko bolj pomembno v prostoru, kjer slovaropisje še ni izkoristilo možnosti, ki jih prinaša digitalni medij (Kosem 2014). Ob tovrstnih analizah pa je seveda treba zagotoviti še številne raziskave slovarske rabe in slovarskih uporabnikov, ki jih v slovenskem prostoru – kljub nekaterim izjemam, ki jih predstavlja pričujoča monografija – še vedno kritično primanjkuje.

IV
Jezikovni viri za slovarski opis sodobne slovenščine

Gradnja referenčnih korpusov na novo: nadgradnja Gigafide

Nataša Logar

Abstract
This paper discusses the expansion of the Gigafida corpus, a reference corpus of Slovenian. In order to become an even better source of language data for a new explanatory monolingual dictionary of contemporary Slovenian, the Gigafida corpus should first of all be supplemented with texts from the period 2010–15 and, if possible, the period 1990–95. In this respect, the issues of copyright and open access to corpus texts are important, as well as issues pertaining to the criteria for the text collection process and the proportions of text types. At the end of the paper, arguments are presented for increasing the number of textbooks in the corpus, and a proposal outlined for a new taxonomy which includes topic/domain categories.

Keywords: reference corpus, Slovenian, dictionary

Ključne besede: referenčni korpus, slovenščina, slovar

1 UVOD

Korpusno jezikoslovje izhaja iz spoznanja, da je jezik v prvi vrsti družbeni pojav, kot tak pa se manifestira izključno v besedilih, ki jih je mogoče opisati in analizirati (Teubert 2005: 108). Središče korpusnega raziskovanja je predvsem performanca (in manj ali pa sploh ne kompetenca) in opazovanje jezika v rabi, ki vodi k teoriji (in ne obratno) (Kennedy 1999: 7; Leech 1992: 107). V tem smislu se korpusno jezikoslovje razlikuje od pristopov k jeziku, ki temeljijo na introspekciji in ne na dokazih (Kennedy 1998: 8). Korpusnih jezikoslovcev ne zanima to, katere besede, strukture ali rabe so v jeziku mogoče, ampak predvsem to, kaj se bo v jezikovni rabi pojavilo kot bolj verjetno, pogosto in tipično ter kaj kot individualno, posebno in enkratno. Korpusi kot vir podatkov za jezikovne opise in utemeljitve so iz tega izhodišča v zadnjih treh desetletjih postali temelj predvsem vsakršne sodobne leksikografije.

»Gradivo za slovar mora biti ustrezno konceptu, zasnovi slovarja. Relevantnost gradiva glede na koncept je temeljnega pomena,« je v razpravnem delu posveta o novem slovarju slovenskega jezika, ki je na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU potekal oktobra 2008, razmišljala Vidovič Muha (Perdih 2009: 35). Ko smo istega leta v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ)[1] pripravljali specifikacijo za zbiranje besedil za korpus, ki bo nadgradil dotedanji referenčni korpus slovenščine FidoPLUS (Arhar Holdt in Gorjanc 2007), smo namen novega korpusa opredelili z naslednjim:

> Znotraj projekta Sporazumevanje v slovenskem jeziku je precej ciljev, katerih uresničitev bo temeljila na novo izdelanem korpusu, med njimi korpusna slovnica /.../ in slogovni priročnik /.../, na korpusu pa bo temeljila tudi celotna leksikalna baza slovenskega jezika, tako v smislu iz korpusa pridobljenih podatkov in njihovih interpretacij kot konkretnih zgledov (*Korpus pisnih besedil: specifikacije /.../*, december 2008: 12).

Korpus Gigafida,[2] ki je bil zaključen leta 2012 (Logar Berginc et al. 2012), je v celoti izpolnil zastavljene cilje, ob njegovi uporabi pri pripravi Leksikalne baze za slovenščino[3] pa smo dobili tudi povratne informacije o njegovih leksikografskih potencialih (Gantar 2009; 2010; 2011). Posledično je bila v *Predlogu za izdelavo* Slovarja sodobnega slovenskega jezika (Krek et al. 2013b) kot izhodišče za pripravo geslovnika novega slovarja navedena prav »frekvenčna lista korpusa Gigafida, v kombinaciji z natančno in razmeroma kompleksno statistično obravnavo podatkov iz korpusa Kres, Gos in drugih baz« (ibid.: 24). Zelo podobno je bilo gradivo za novi slovar opredeljeno tudi v Gliha Komac et

1 http://projekt.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx (dostop 6. 7. 2015).
2 http://www.gigafida.net (dostop 6. 7. 2015).
3 http://www.slovenscina.eu/spletni-slovar/leksikalna-baza (dostop 6. 7. 2015).

al. (2015: 4): »Gradivo za izdelavo geslovnika in redakcijsko obdelavo osrednjih delov posameznih slovarskih sestavkov /.../ so korpusni viri, predvsem Gigafida, Kres, Nova beseda in deloma Gos.« Lahko torej še enkrat zapišemo, da so si bili ključni slovenski leksikografi v letu 2015 o vlogi Gigafide in Kresa pri prihodnjem velikem slovenskem slovarskem podvigu enotni: oba korpusa sta ustrezna gradivna osnova za prikaz leksikalne podobe javne pisne slovenščine zadnjih 20 let (Logar et al. 2015; prim. tudi Logar 2014: 10) – seveda z dodatkom, da je potrebna njuna nadgradnja.

Nadgradnja Gigafide (in Kresa)[4] je najprej potrebna zato, ker so bila zadnja besedila iz tiska, ki so vključena vanjo, pridobljena 29. 5. 2010, precej ozko usmerjeno zbirana besedila z interneta pa obsegajo zgolj obdobje od 1. 4. 2010 do 11. 4. 2011 (Logar Berginc et al. 2012: 43). V času priprave tega prispevka torej besedil iz knjig, revij in časopisov, ki bi bila mlajša od petih let, v Gigafidi ni. Drugi, morda še pomembnejši razlog za posodobitev pa je v zelo spremenjenih in razširjenih možnostih dostopa do javne besede, ki so podobo širokim množicam namenjene slovenščine močno spremenile, preobrazile marsikateri žanr, ki je bil do tedaj vezan le na tisk, in z njim povezano urejanje ter prinesle nove, tudi po jeziku specifične vrste pisnih besedil. In kot smo zapisali že v Logar in Ljubešić (2013: 104):

> V zagovor nujnosti gradnje korpusov – takrat sicer korpusov *govorjenih* besedil – sta Stabej in Vitez leta 2000 zapisala: 'dejstvo je, da je analitična slika nekega jezika, ki elemente zajema samo iz pisnih besedil, izrazito delna in nepopolna' (79). In dalje še: 'če je idealni cilj korpusno podprtega jezikoslovja spoznavanje jezika, kot je izpričan v vseh razsežnostih sporazumevanja, je samo pisni korpus premalo' (80). Navedeno je mogoče oz. celo nujno prenesti na besedila, ki jih desetletje pozneje pišemo za 'nove medije' in beremo na njih. Njihova vnaprejšnja opustitev iz korpusov, ki so osnova za jeziko(slo)vne opise jezika v vseh razsežnostih sporazumevanja in utemeljitve zanje, bi pomenila diskvalifikacijo pomembnega dela jezika.

Krek (11. 11. 2013) je na zaključni konferenci projekta SSJ poudaril, da se v času priprave specifikacij za Gigafido še nismo zavedali, do kako velikega porasta uporabe družabnih omrežij in z internetom povezanih mobilnih naprav bo prišlo po letu 2008, pa hkrati npr. tudi dejstva, da bo v istem času močno upadlo branje tiskanih časopisov. V luči te nove družbene realnosti, ki močno vpliva na jezik in z njim povezane opise ter vire in tehnologije, je zato treba gradnjo referenčnih korpusov premisliti na novo, pri čemer je smiselno izhajati iz dobrih preteklih praks doma in v tujini ter načrtovati popravke tam, kjer je korpusna analiza že utemeljeno izpostavila pomanjkljivosti.

4 Kadar je smiselno, imamo v nadaljevanju v mislih oba.

V nadaljevanju poglavja bomo zato razmišljali o tistih segmentih nadgradnje Gigafide, ki bi ta korpus kot gradivni vir za razlagalni enojezični slovar sodobne slovenščine naredili še ustreznejši in relevantnejši, s tem da tukajšnje razpravljanje v delih, ki zahtevajo obširnejšo obravnavo (spletna besedila ter označitev in zapis), dopolnjujeta naslednja dva prispevka v tem poglavju.

2 SODOBNA SLOVENŠČINA

2.1 Začetek zajema besedil: leto 1990

Sodobnost jezika je relativen pojem, in če želimo z njim opredeliti časovno razsežnost besedil, ki jih zajema korpus, postane ta pojem nujno tudi dogovorni. Na dogovor o »sodobnosti« vplivajo tako zunaj- kot znotrajjezikovni dejavniki, merodajne za določitev letnic pa so predvsem večje spremembe obojih. V korpusno-slovarski praksi se v tem smislu kot razlog za začetno (večinoma na desetletje zaokroženo) letnico zajema besedil največkrat navajajo:

- čas izida predhodnega splošnega slovarja,
- večje spremembe v družbenopolitični ureditvi, ki prinesejo večje spremembe v leksiki, in
- praktični razlogi, npr. obstoj elektronskih arhivov pri besedilodajalcih ali odstop besedil.

Če si za izhodišče najprej ogledamo stanje sodobnih korpusov in splošnih slovarjev češkega ter slovaškega jezika, ki sta po letu 1989 zaradi družbenih, političnih in gospodarskih dogodkov svojo poimenovalno podobo (pa sicer tudi status jezika in z njim povezane govorne položaje) spremenila (oz. razširila) podobno kot slovenščina,[5] ugotovimo naslednje:

a) Avtorji uravnoteženega referenčnega korpusa češkega jezika, ki ga pripravljajo na Inštitutu za Češki narodni korpus na Filozofski fakulteti v Pragi, so o prvem korpusu, ki je bil izdelan leta 2000 (SYN2000;[6] sledila sta mu nato še SYN2005 in SYN2010), zapisali, da gre za »sinhroni korpus, ki vsebuje sodobno češčino, torej predvsem besedila, ki so nastala v letih 1990–1999«, ter da je bilo za publicistiko in strokovna besedila leto 1990 izbrano kot »naravni mejnik sinhronije«. Enako je veljalo tudi za jedrni del leposlovja, s to izjemo, da so bila v korpus vključena tudi leposlovna dela starejšega datuma, to je tista, ki se še vedno ponatiskujejo in torej vplivajo na sodobno češčino (pri čemer se je moral njihov avtor

5 Prim. tudi prispevek o latvijščini: Migla in Zuicena (2014).
6 https://ucnk.ff.cuni.cz/english/syn2000.php (dostop 6. 7. 2015).

roditi po letu 1880; taka so npr. dela K. Čapka in J. Haška).[7] Do danes najsodobnejši slovar češčine Slovník spisovného jazyka českého je sicer precej starejšega datuma – v štirih zvezkih je izšel v letih 1960–1971, Inštitut za češki jezik Češke akademije znanosti pa ga je leta 2011 objavil tudi na spletu.[8] Na Inštitutu za češki jezik v tem času pripravljajo nov slovar z naslovom Akademski slovar sodobne češčine (Akademický slovník současné češtiny), vendar iz trenutno zelo skopih objav ni razvidna njegova korpusna zasnova.[9]

b) Tudi na Znanstvenojezikoslovnem inštitutu L'udevíta Štúra Slovaške akademije znanosti pravkar pripravljajo nov slovar, ki nosi naslov Slovar sodobnega slovaškega jezika (Slovník súčasného slovenského jazyka). Izšla sta že dva zvezka: prvi leta 2006 (A–G), drugi leta 2011 (H–L). Načrtovan je kot slovar velikega obsega s približno 220.000 iztočnicami, sicer pa je bil njegov predhodnik, tj. Slovar slovaškega jezika (Slovník slovenského jazyka), izdan že štiri desetletja prej, v letih 1959–1968 (Buzássyová 2009: 119). Primarno gradivo novega slovarja je leksikografska kartoteka s petimi milijoni listkov in Slovaški narodni korpus,[10] ki ga gradijo od leta 2002 (ibid.: 124), vsebuje pa besedila od leta 1955 dalje (Šimková in Garabík 2014). Buzássyová, ki je glavna urednica slovarja, je leta 2008 o sodobnosti slovarja in njegovega gradiva povedala še naslednje (Perdih 2009: 52):

> V /slovaški/ teoriji se za sodobni jezik razumejo tudi 40. leta 20. stoletja, ko se je Češkoslovaška prvič razdelila. Slovaška država in jezik te družbe je /takrat/ prvič prevzel vse funkcije, na primer jezik umetnosti, leposlovja, govorjeni jezik, jezik administracije, izrazito tudi strokovni jezik, vendar ne izhajamo od 40. let, ker to ne bi bilo realno. /.../ Izhajamo iz 2. svetovne vojne, kar je do 60. let zajel tudi predhodni slovar.

Odločitve in razlogi čeških ter slovaških korpusnih jezikoslovcev in leksikografov potrjujejo zelo podobne utemeljitve izpred desetih let o sodobnosti besedil v prvem slovenskem referenčnem korpusu FIDA, nadgradnji katerega sta bili potem FidaPLUS in današnja Gigafida. Prim. Gorjanc (2005: 47–48):

> Korpus FIDA skuša posredovati vsestranske informacije o sodobnem slovenskem jeziku, torej z besedili skuša zajeti čim bolj celovito podobo današnje slovenščine /.../. Korpus FIDA je *sinhroni korpus*; vanj so vključena besedila po l. 1990. /.../ /P/rvotna ideja o vključevanju besedil po l. 1980 je bila že na samem začetku gradnje korpusa spremenjena iz dveh ključnih razlogov. Prvi, povsem pragmatični, je vezan na poizvedovanje po dostopnih besedilih v elektronski obliki; pokazalo se je, da se je kultura elektronskih arhivov začela oblikovati šele v drugi polovici

7 http://wiki.korpus.cz/doku.php/cnk:syn2000 (dostop 6. 7. 2015).
8 http://ssjc.ujc.cas.cz/ (dostop 6. 7. 2015).
9 http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/akademicky-slovnik-soucasne-cestiny.html (dostop 6. 7. 2015).
10 http://korpus.juls.savba.sk/ (dostop 6. 7. 2015).

> devetdesetih let, tako da bi tovrstna besedila morali pred vključitvijo v korpus digitalizirati. Drugi je povezan s kartotečno zbirko Inštituta za slovenski jezika Frana Ramovša ZRC SAZU, ki nekako do tega časa zagotavlja vsaj osnovno informacijo o stanju jezika še v osemdesetih letih prejšnjega stoletja.

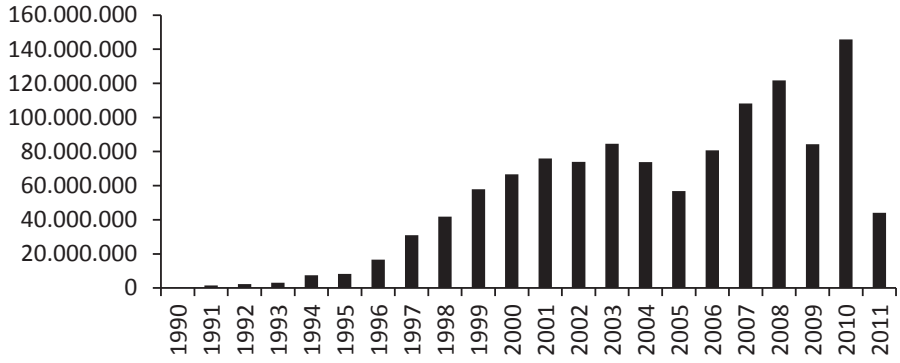
Ter v Logar Berginc et al. (2012: 127):

> V procesu definiranja časa zajema besedil je pri sestavljalcih /korpusa FIDA/ prevladalo mnenje, da je menjava političnega sistema v Sloveniji na rabo jezika vplivala dovolj, da je to letnico mogoče vzeti kot izhodišče za pojem 'sinhronosti' korpusa. /.../ Korpus je torej zajemal desetletno obdobje od leta 1991 do 2000, z nekaj besedili iz let 1989/90.

Če povzamemo: začetek zajema besedil za Gigafido in njeno prihodnjo nadgradnjo za potrebe slovarja postavljamo v leto 1990 in to utemeljujemo z naslednjim: (a) časom izida zadnje knjige Slovarja slovenskega knjižnega jezika (1970–1991; SSKJ), (b) družbenopolitičnimi spremembami konec 80. let prejšnjega stoletja, zlasti pa po osamosvojitvenem letu 1991, ki so bistveno vplivale na izrazno podobo današnje slovenščine, ter (c) praktičnim razlogom, tj. obstojem elektronskih arhivov pri založbah in drugih besedilodajalcih.

2.2 Besedila po letu 2010 in iz prve polovice 90. let prejšnjega stoletja

Časovno obdobje, ki ga pokrivajo besedila v Gigafidi, se začenja v že omenjenem letu 1990 in zaključuje v prav tako že pojasnjenem letu 2010 (tisk) oz. 2011 (internet). Število besed po letih kaže Slika 1.



Slika 1: Število besed po letih v Gigafidi. Vir: Logar Berginc et al. (2012: 36).

Izkušnje kažejo, da se besedila iz tiska pridobijo predvsem za leto in nekaj let nazaj, manj pa za tekoče leto zbiranja, zato sta upada v letu 2005 in 2009 pričakovana. Ti dve leti je mogoče še dopolniti, če odlog do naslednjega zbiranja ne bo trajal predolgo. Nadgradnja s spletnimi besedili je sprotna in vezana na čas trajanja projekta, nato pa se prekine. V tem smislu pajkanja za obdobje 2012–2015, ki bi ga vodil referenčnokorpusni namen, ne moremo več v celoti nadomestiti.[11] To dejstvo, pa tudi vrzeli v naboru tiskanih besedil, ki so posledica predolgih obdobij neposodabljanja korpusov, vsekakor govorijo v prid dolgoročnejše infrastrukturne rešitve, kakršna je npr. Spletni arhiv Narodne in univerzitetne knjižnice[12] ali urejeno dolgoročno financiranje infrastrukturnih programov Centra za jezikovne vire in tehnologije UL.[13]

Hkrati je v korpusu zelo malo besedil, ki bi omogočala natančnejši vpogled v leksikalni nabor slovenščine prve polovice 90. let 20. stoletja. Za sedemletno obdobje 1990–1996 Gigafida vsebuje sicer na prvi pogled obsežnih 22 milijonov besed, ki pa v celotnem korpusu predstavljajo manj kot 2-odstotni delež. Če bi bil prihodnji projekt nadgradnje Gigafide finančno in časovno dovolj širok, da bi omogočal tudi digitalizacijo izbranih besedil iz tega obdobja, bi bilo o tovrstni dopolnitvi korpusa vsekakor vredno razmisliti.

3 SLOVENŠČINA V SPLOŠNI PISNI RABI

3.1 Ustreznost korpusa za slovarske potrebe in namene

O tem, kaj je usmerjalo zbiranje besedil ob vsakokratnem korpusu iz »serije FIDA«, s(m)o pisali že večkrat. V grobem gre za naslednje:

a) Namen: Korpusi FIDA, FidaPLUS in Gigafida so bili zgrajeni zato, da bi prikazovali celovito podobo slovenskega jezika, kot se kaže v javno objavljenih pisnih besedilih. V tem smislu je torej Gigafida kot zadnja nadgradnja namenjena različnim jezikoslovnim raziskavam, v ospredju (kot tudi sicer to velja za splošne ali referenčne korpuse) pa je njena uporabnost za leksikološke in leksikografske namene (Gorjanc et al. 2005; Gantar 2009; Kosem et al. 2012).

b) Merila zbiranja besedil, vsebina in dokumentiranost: Tako Gigafida kot njeni predhodnici FIDA in FidaPLUS sta imeli jasno razvidna merila zbiranja besedil, vsa ta merila, kot tudi posamezne ločene odločitve in uspešnost zbiranja v skladu z njimi so popisani tudi v literaturi (Erjavec

11 Lahko bi si sicer delno pomagali s spletnim korpusom slovenščine slWaC$_2$ (Erjavec in Ljubešić 2014), a zbiranje spletnih besedil za ta vir ni bilo usmerjevano in kontrolirano, pa tudi časovno predvidljivo ter enakomerno na način, ki bi bil zaželen za nadgradnjo Gigafide (več gl. v poglavju IV).

12 http://arhiv.nuk.uni-lj.si/ (dostop 6. 7. 2015).

13 http://www.cjvt.si/ (dostop 6. 7. 2015).

1998; Erjavec, Gorjanc in Stabej 1998; Gorjanc 1999; Gorjanc 2000; Gorjanc 2005; Arhar Holdt in Gorjanc 2007; Romih 1998; Stabej 1998; Železnikar 1998; Logar Berginc in Šuster 2009; Logar Berginc et al. 2012: 119–136).

c) »Lovljenje« splošne rabe: Merila zbiranja besedil so že od korpusa FIDA dalje izhajala tako iz recepcije kot produkcije (gl. literaturo v prejšnjem odstavku); v zvezi z recepcijo – kolikor se je dalo – skozi sito širše vplivnosti. Pri tem smo upoštevali objektivne podatke o branosti (Logar Berginc et al. 2012: 14–25, 46–48) na podlagi: Nacionalne raziskave branosti (časopisi, revije); izposoje v knjižnicah, knjižnih nagrad, naklade, obiskanosti spletnih strani ipd. Zbiranje specializiranih besedil (znanstvenih) smo pri tretjem zbiranju opustili, zato jih je v Gigafidi malo. V kolikšni meri Gigafida dejansko prikazuje splošno pisno rabo, je seveda težko oceniti, nikoli pa niso zbiralci odstopili od osrednje težnje, ki je bila: táko rabo vendarle skušati ujeti (kot rečeno, z upoštevanjem recepcije in produkcije).

Po številu besed v Gigafidi prevladuje periodični tisk s 77 %. Ker smo se vnaprej zavedali, da bo rezultat verjetno takšen, smo v projektu SSJ iz Gigafide vzorčili še Kres (Erjavec in Logar Berginc 2012).

Gigafida je torej velik ter po času, zvrsteh, avtorjih, temah idr. raznolik korpus. Krek in Kosem (21. 9. 2013) sta v zvezi s tem zapisala: »/Č/im več govorcev dejansko bere določena besedila (ne glede na njihovo 'slogovno šibkost'), tem večji vpliv imajo ta na njihov jezik in toliko bolj so pomembna za leksikografsko obravnavo, ki v konsistentno zasnovanem procesu vsebino slovarske baze opremi z relevantnimi informacijami za različne tipe uporabnikov.« Na podlagi povedanega se zdi tudi pri nadgradnji Gigafide in Kresa, ki (ali če) bosta pri pripravi novega splošnega slovarja glavna vira podatkov o podobi sodobne javne pisne slovenščine, smiselno še naprej slediti načelu večje sporočanjsko-vplivanjske vloge besedil z manjšo (ali celo nikakršno) vlogo ozko specializiranih znanstvenih besedil, zasebnih besedil in vseh drugih besedil, ki imajo majhno recepcijo (gl. tudi razdelek 6.1 in poglavje VII v tej monografiji).

3.2 Vprašanje »metakorpusa«

Oba uvodna navedka iz dveh predlogov prihodnjega slovarja slovenščine (Krek et al. 2013b in Gliha Komac et al. 2015) kot vir za geslovnik in redakcijo slovarskih sestavkov ob Gigafidi navajata še kombiniranje s Kresom, Gosom, Novo besedo ter drugimi bazami podatkov. V zadnjem desetletju je v slovenskem prostoru nastal dokaj obširen nabor različnih korpusov (prim. npr.

Erjavec 2013),[14] zato se samo po sebi odpira vprašanje povezljivosti vseh v enega (prim. tudi Gorjanc 2009: 47) in nato uporaba le-tega v slovarske namene. Ali kot smo zapisali v Logar et al. (2015): »Za prihodnje slovarsko delo /.../ ni pomembno le vprašanje, kateri korpusi *bodo* slovarsko gradivo in zakaj, temveč tudi vprašanje, kateri korpusi *ne* bodo slovarsko gradivo in zakaj ne.«

Tu zagovarjamo odločitev, da mora biti korpus, ki bo glavno gradivo za splošni slovar, že *narejen s takim namenom*, da mora biti natančno *dokumentiran* ter po *vsebini in zgradbi razviden.* Zgolj na ta način bo korpus kot vzorec sploh dopuščal posploševanja, ki bodo nato izšla kot splošnojezikovni opis in predpis. Ob glavnem slovarskem viru (v našem primeru kot takega razumemo Gigafido skupaj z njeno izvedenko Kresom) so seveda mogoča tudi kombiniranja z drugimi korpusnimi viri in bazami podatkov (taka je npr. leksikografska praksa pri trenutno nastajajočem Velikem slovarju poljskega jezika, prim. Żmigrodzki 2014: 2), vendar pa vedno z zelo jasno razvidnim namenom ter na način, ki bo uporabnikom slovarja pojasnjen, v uredniškem postopku pa natančno predpisan.

4 BESEDILNA AVTORSKOPRAVNA RAZMERJA IN ODPRTI DOSTOP

Korpusi FIDA, FidaPLUS in Gigafida so imeli pravna razmerja z besedilodajalci urejena tako, da je bilo mogoče korpus objaviti javno in v prostem dostopu. Pri tem je bil bistven pogodbeni prenos materialnih avtorskih pravic nad besedilom na način, kot ga določa 22. člen Zakona o avtorskih in sorodnih pravicah (ZASP 2007). Ker je šlo pri tem za dostop do besedila v digitalni obliki, je imetnik pravic na pripravljalce korpusa prenašal tudi pravici elektronskega reproduciranja, kot je določeno v prvem odstavku 23. členu ZASP, in predelave, kot je določeno v 33. členu ZASP:

> 22. člen:
> (1) Pravica reproduciranja je izključna pravica, da se delo fiksira na materialnem nosilcu ali drugem primerku, in sicer neposredno ali posredno, začasno ali trajno, delno ali v celoti ter s kakršnimkoli sredstvom ali v katerikoli obliki.
>
> 23. člen:
> (1) Pravica predelave je izključna pravica, da se neko prvotno delo prevede, odrsko priredi, glasbeno aranžira, spremeni ali kako drugače predela.
> (2) Pravica iz prejšnjega odstavka se nanaša tudi na primere, ko se prvotno delo v nespremenjeni obliki vključi ali vgradi v novo delo.
> (3) Avtor prvotnega dela obdrži izključno pravico do uporabe svojega dela v predelani obliki, če ni s tem zakonom ali s pogodbo drugače določeno.

14 http://nl.ijs.si (dostop 6. 7. 2015).

V pogodbi med besedilodajalci in pripravljalci korpusa Gigafida je bil tudi člen, po katerem je imetnik pravic dovolil, da se do 10 % besedila uporabi na način, kot ga določa licenca *Creative Commons: priznanje avtorstva + nekomercialno + deljenje pod enakimi pogoji*, bolj znana pod oznako CC BY-NC-SA.[15] Ta člen je omogočil izdelavo korpusov ccGigafida (v obsegu 100 milijonov besed) in ccKres (10 milijonov besed), ki sta prosto dostopna v obliki baze podatkov.[16]

Odprti dostop do raziskovalnih podatkov iz javno financiranih projektov so s podpisom Deklaracije o dostopu do raziskovalnih podatkov iz javnega financiranja (angl. *Declaration on Acces to Research Data from Public Funding*, 2004) podprle vse članice organizacije OECD, izrecno ob pridružitvi leta 2010 tudi Slovenija (prim. tudi smernice in načela dostopa do raziskovalnih podatkov iz javnega financiranja iste organizacije – *OECD Principles and Guidelines for Access to Research Data from Public Funding*).[17] Pobudi so se s strateškimi dokumenti, poročili in zavezami pridružili tudi Evropska komisija, Evropski znanstveni svet, Evropska federacija Akademij znanosti ALLEA in drugi, zlasti zavezujoče pa je v tem okviru priporočilo Evropske komisije o dostopu do znanstvenih informacij in njihovem arhiviranju iz leta 2012.[18] Ta države članice EU med drugi in poziva, naj bo dostop do »objav, ki so rezultat javno financiranih raziskav, odprt čim prej, po možnosti takoj, v vsakem primeru pa najpozneje šest mesecev po datumu objave, za družbene znanosti in humanistične vede pa dvanajst mesecev« (L194/41).[19] V zaključnem poročilu CRP-projekta Odprti podatki – Priprava akcijskega načrta za vzpostavitev sistema odprtega dostopa do podatkov iz javno financiranih raziskav v Sloveniji (2010–2013) so raziskovalci poudarili, da so odprti raziskovalni podatki

> skupna odgovornost vseh akterjev v znanosti, ki ne more biti prepuščena samo enemu segmentu, npr. etičnim načelom, pač pa zahteva jasno opredeljene obveznosti tako posameznih raziskovalcev, njihovih ustanov in vodstev, strokovnih in znanstvenih društev ter drugih predstavnikov znanstvene skupnosti, izvajalcev s podatki povezanih storitev, založnikov (Štebe et al. 2013: XVI).

Pri prihodnji gradnji referenčnega korpusa slovenščine se bo tako treba ponovno zavezati tej odgovornosti in korpus pripraviti ne le za rabo v konkordančniku, temveč znova vsaj v omejenem obsegu tudi v obliki »CC«,[20] ki bo omogočala domačim in tujim raziskovalcem »razvoj kakovostnih, robustnih in praktično uporabnih orodij za obdelavo naravnega /v našem primeru slovenskega/ jezika«

15 https://creativecommons.org/licenses/by-nc-sa/2.5/si/legalcode (dostop 6. 7. 2015).
16 http://hdl.handle.net/11356/1035 in http://hdl.handle.net/11356/1034 (dostop 6. 7. 2015).
17 http://www.oecd.org/sti/sci-tech/38500813.pdf (dostop 6. 7. 2015).
18 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF (dostop 6. 7. 2015).
19 Več o odprtem dostopu gl. na http://www.openaccess.si/ (dostop 6. 7. 2015).
20 V omejenem obsegu zato, ker od besedilodajalcev (z redkimi izjemami) ni mogoče pričakovati strinjanja z uporabo celotnega besedila pod licenco Creative Commons.

(Erjavec 2009: 115; Erjavec 2014). Da so taka orodja za slovenščino nujno potrebna, je bilo opozorjeno že večkrat (npr. Krek 2012).

5 SORODNI KORPUSI V SODOBNI TUJI LEKSIKOGRAFSKI PRAKSI

Tabela 1 prikazuje seznam trenutno nastajajočih ali pred kratkim nastalih splošnih slovarjev finskega, estonskega, latvijskega, poljskega, češkega, slovaškega, nizozemskega, nemškega in angleškega jezika skupaj z zgradbo korpusa, ki je (bil) slovarska podlaga (če tak korpus seveda obstaja).[21]

Tabela 1: Seznam slovarjev devetih tujih jezikov skupaj z obsegom in zgradbo korpusov, iz katerih so nastali oz. še nastajajo. Vir: Dopolnjeno in posodobljeno glede na Logar (2014).

Jezik, slovar, korpus[22]	Obseg korpusa	Zgradba korpusa
FINŠČINA Novi slovar sodobne finščine / Kielitoimiston sanakirja	Slovar ni korpusno zasnovan (Heinonen 2014).	/
ESTONŠČINA Osnovni slovar estonščine / The Basic Estonian Dictionary (spletna izdaja je v pripravi; Kallas et al. 2014) Uravnoteženi korpus estonščine / The Balanced Corpus of Estonian (http://www.cl.ut.ee/korpused/grammatikakorpus/)	15 mio	• časopisi in revije: 33 % • leposlovje: 33 % • znanstvena besedila: 33 %
LATVIJŠČINA Slovar sodobnega latvijskega jezika / Mūsdienu latviešu valodas vārdnīca (www.tezaurs.lv/mlvv) Uravnoteženi korpus sodobne latvijščine / Līdzsvarots mūsdienu latviešu valodas tekstu korpuss (www.korpuss.lv)	4,5 mio	• časopisi in revije: 55 % • leposlovje: 20 % • znanstvena besedila: 10 % • pravna besedila: 8 % • drugo: 5 % • zapisi parlamentarnih sej: 2 %

21 Če za posamezen jezik nastaja več takih splošnih slovarjev, smo izbrali tistega, ki je zasnovan tudi za objavo na spletu; če je takih slovarjev več (angleščina), pa je bil izbor naključen.

22 Vse v tabeli navedene spletne strani smo si zadnjič ogledali 18. 5. 2015.

Jezik, slovar, korpus[22]	Obseg korpusa	Zgradba korpusa
POLJŠČINA Veliki slovar poljskega jezika / Wielky słownik języka polskiego (http://www.wsjp.pl/) Nacionalni korpus poljskega jezika / Narodowy korpus języka polskiego (http://nkjp.pl/)	(načrtovano) 1,5 mld (Górski in Łazinski 2012: 33)	• časopisi, revije in sporočila za javnost: 50 % • leposlovje: 16 % • govorjena besedila: 10 % • stvarna besedila: 11 % • spletna besedila: 7 % • didaktična besedila: 2 % • drugo: 3 % • neuvrščeno: 1 %
ČEŠČINA Akademski slovar sodobne češčine / Akademický slovník současné češtiny (http://www.ujc.cas.cz/zakladni-informace/oddeleni/oddeleni-soucasne-lexikologie-a-lexikografie/akademicky-slovnik-soucasne-cestiny.html)	Podatek o korpusni gradivni zasnovi ni naveden oz. razviden.	/
SLOVAŠČINA Slovar sodobnega slovaškega jezika / Slovník súčasného slovenského jazyka (http://slovniky.juls.savba.sk/) Slovaški nacionalni korpus / Slovenský národný korpus (2013) (http://korpus.juls.savba.sk/stats.html)	829 mio	• časopisi in revije: 69 % • stvarna besedila: 15 % • leposlovje: 14 % • drugo: 2 %
NIZOZEMŠČINA Splošni nizozemski slovar / Algemeen Nederlands Woordenboek (http://anw.inl.nl/search) ANW-korpus / Algemeen Nederlands Woordenboek (ANW) Corpus (http://anw.inl.nl/show?page=help_anwcorpus)	102,5 mio	• časopisi: 40 % • spletna besedila: 30 % • leposlovje: 20 % • časopisi, revije in novičarski portali – izbor zaradi neologizmov: 5 % • starejša besedila, 1970–2000: 5 %

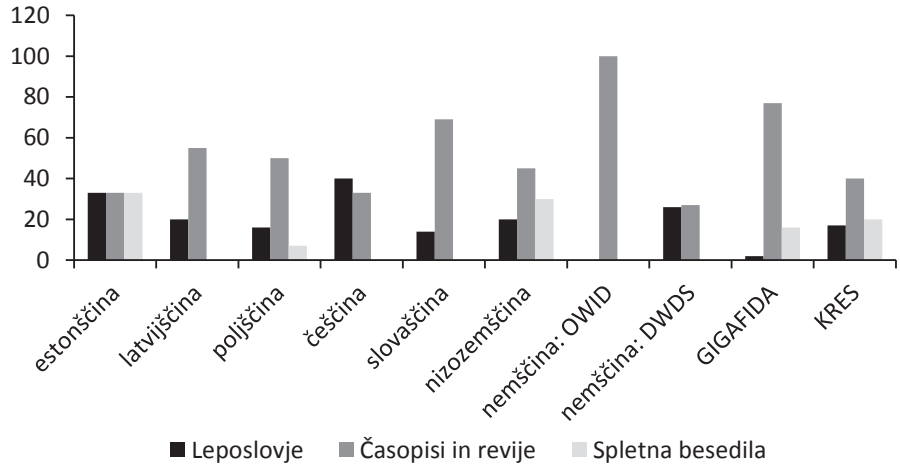
NEMŠČINA a) projekt OWID Inštituta za nemški jezik v Mannheimu (http://www1.ids-mannheim.de/lexik/owid.html) Elexiko (http://www.owid.de/wb/elexiko/start.html) Elexiko-korpus (http://www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html)	2,7 mld	• časopisi in revije: 100 %
b) DWDS: Digitalni slovar nemškega jezika / Das Digitale Wörterbuch der deutschen Sprache (http://www.dwds.de/) Kernkorpus[23] (http://www.dwds.de/ressourcen/korpora/)	122 mio	• leposlovje: 26 % • stvarna besedila: 22 % • znanstvena besedila: 25 % • časopisi in revije: 27 %
ANGLEŠČINA Oxford Dictionaries (http://www.oed.com/) Oxford English Corpus (http://www.oxforddictionaries.com/words/the-oxford-english-corpus)	2,5 mld	• besedila s spleta: skoraj 100 % (romani, nespecializirane in specializirane revije, časopisi, blogovski zapisi, e-pošta, družbena omrežja ipd.)

Iz tabele je razvidno, da so korpusi, ki so gradivo za trenutno aktualne in primerjalno zanimive slovarje sedmih tujih jezikov (če finskega in češkega odštejemo), po svoji zgradbi zelo različni. Če se v nadaljnji primerjavi omejimo le na tri ključne, pri Gigafidi v kritikah najbolj izpostavljene kategorije, tj. na leposlovje, publicistiko in spletna besedila, dobimo podatke, ki jih kažeta Tabela 2 in Slika 2 (izpuščamo tudi angleški korpus, katerega notranja členitev ni javno objavljena, dodajamo pa podatke za primerjalno zanimiv češki korpus SYN2010).

23 Slovar sicer nastaja iz 15 korpusov, Kernkorpus kot uravnoteženi in referenčni korpus je njegovo osrednje gradivo.

Tabela 2: Zgradba korpusov sedmih tujih jezikov ter Gigafide in Kresa (v %) pri kategorijah leposlovje, časopisi in revije ter spletna besedila. Vir: Dopolnjeno in posodobljeno glede na Logar (2014).

	Leposlovje	Časopisi in revije	Spletna besedila
estonščina	33	33	33
latvijščina	20	55	0
poljščina	16	50	7
češčina	40	33	0
slovaščina	14	69	0
nizozemščina	20	45	30
nemščina: OWID	0	100	0
nemščina: DWDS	26	27	0
GIGAFIDA	2	77	16
KRES	17	40	20



Slika 2: Zgradba korpusov sedmih tujih jezikov ter Gigafide in Kresa (v %) pri kategorijah leposlovje, časopisi in revije ter spletna besedila. Vir: Dopolnjeno in posodobljeno glede na Logar (2014).

V Tabeli 2 in na Sliki 2 lahko ugotovimo naslednje: v povprečju največ besedil v korpuse prihaja iz časopisov in revij; Gigafida izstopa navzdol pri leposlovju, po deležu publicistike pa sodi v vrh, čeprav jo tu presega nemški korpus projekta OWID, blizu pa ji je tudi korpus slovaščine. Po deležu spletnih besedil je Gigafida primerjalno približno na sredini. Kres je glede na druge korpuse povprečen.

6 USMERJENO ZBIRANJE BESEDIL ZA POTREBE SLOVARJA

6.1 Specializirana leksika

Ledinek (2014: 2) je povzela bistvena vprašanja, povezana z vključitvijo specializirane leksike v splošne slovarje, z naslednjim:

> Vprašanja, kaj je terminologija v konkretnem enojezičnem razlagalnem slovarju srednjega obsega, kolikšen bo v slovarju njen predpostavljeni delež, katera strokovna področja bodo (v večji meri in sistematično) vključena in kakšen bo način terminološkega kvalificiranja (izhodiščno) terminološke leksike, so temeljna vprašanja slovarskega koncepta.

Strinjati se je mogoče, da ni nobenega dvoma o tem, ali specializirano leksiko vključiti v splošni slovar s približno 100.000 iztočnicami ali ne, vprašanje pa je, kaj sploh je specializirana leksika z vidika splošnega slovarja ter kako jo skupaj z njenim tipičnim besedilnim okoljem vanj vključiti (več gl. v poglavju VII). Poleg tega je vprašljivo tudi vnaprejšnje določanje deleža specializirane leksike. Da pa bi bil kakršenkoli nabor in izbor sploh mogoč, je treba korpus, iz katerega bo nastal slovar, pripraviti tako, da bo čim bolje odslikaval stanje terminološke – oz. determinologizirane – leksike v splošnem jeziku. Če tu pustimo ob strani dejstvo, da tako leksiko kažejo že v korpus vključeni časopisi in novičarski spletni portali, je za dosego tega cilja v Gigafidi pri dopolnitvi z novimi besedili smiselno slediti dvema načeloma:

a) načelu *nevključevanja* področno specializiranih besedil (znanstvenih revij in monografij, doktorskih disertacij, prispevkov na znanstvenih konferencah ipd. – torej ravno tistih, ki so najbolj zanimiva za korpuse strokovnih besedil; prim. Logar 2013: 47–52) ter hkrati

b) načelu *vključevanja* poljudnostrokovnih del in učbenikov do vključno ravni srednje šole.

Zapisali smo že, da smo se pri zadnjem zbiranju znanstvenim besedilom izognili, medtem ko je bila velika pozornost v celotnem obdobju zbiranja po letu 1997 namenjena pridobivanju poljudnostrokovnih knjig (priročnikov, vodnikov ipd.) z različnih področij človekovega življenja ter revij, ki strokovna področja upovedujejo na laikom (pogosto mlajšim bralcem) razumljiv način. Gigafida tako vsebuje skoraj 900 priročniških del 84 različnih založnikov, izmed revij pa jih vsaj 50 ustreza opisu poljudne strokovnosti (npr. za avtomobilizem: *Avto foto market*, *Avto magazin*, *Avtokatalog*, *Motorevija*, *Motokatalog* in *Mobil*; za računalništvo: *Connect*, *Joker*, *Moj mikro*, *Monitor*, *PC & mediji* in *Računalniške novice*). V tem smislu je mogoče tudi pri novem zbiranju izhajati iz predhodnih dobrih praks in izkušenj. Drugače pa je

pri učbenikih, katalogih znanj in didaktičnih pripomočkih, pri katerih bi moralo biti novo zbiranje bolj načrtno. Gigafida sicer vsebuje 103 taka dela, ki jih je odstopilo pet založb: Zavod RS za šolstvo, RIC Državni izpitni center, Rokus Klett, DZS in Ataja, a pregled vključenih učbenikov oz. delovnih zvezkov kaže, da so področja obveznega osnovnošolskega programa z njimi zajeta zelo neenakomerno:

- matematika (6 učbenikov oz. delovnih zvezkov)
- slovenščina (13)
- angleščina (1)
- zgodovina (8)
- biologija (7)
- spoznavanje okolja (2)
- fizika (1)
- kemija (4)
- družba (4)
- naravoslovje (1)
- naravoslovje in tehnika (1)
- likovna umetnost (1)
- glasbena umetnost (8)
- šport (1)
- gospodinjstvo (3)

Že na prvi pogled je torej razvidno, da je v Gigafidi obvezni šolski program z učbeniki pokrit slabo, za programe srednjih šol je del še veliko manj. Glede na Predmetnik osnovne šole Ministrstva za izobraževanje, znanost in šport RS[24] povsem manjkajo še učbeniki za geografijo, domovinsko in državljansko kulturo ter tehniko in tehnologijo. S tega vidika je treba korpus dopolniti, najbolje s težnjo po zajemu učbenikov in delovnih zvezkov ter sorodnih učencem ter dijakom namenjenih besedil vseh šolskih predmetov splošnih in poklicnih programov (osnovne šole, gimnazije, poklicne srednje šole). Poleg tega bi bilo dobro pridobiti tudi podatke o učbenikih in podobnih gradivih za obšolske interesne dejavnosti, zlasti tiste, ki so množično obiskane, in skušati pridobiti tudi ta gradiva. Na ta način bi nadgrajena Gigafida – ob predpostavki naklonjenih besedilodajalcev – ustrezno zajela terminologijo, s katero se v okviru predterciarnega institucionalnega izobraževanja sreča skoraj vsa učeča se populacija, iz takega korpusa nastali nabor terminov pa bi bil v bolj celostnem obsegu na razpolago za nadaljnji, s konceptom slovarja usklajen leksikografsko-terminografski postopek.

24 http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/podrocje/os/devetletka/predmetniki/Pred_14_OS_4_12.pdf (dostop 6. 7. 2015).

6.2 Tematska pokritost

Zbiranje besedil za referenčne korpuse usmerja več meril, med katerimi je tudi raznovrstnost besedilnih tem. Pri zbiranju besedil za Gigafido smo izhajali iz naslednjega seznama (Logar Berginc et al. 2012: 15):

- aktualni dogodki
- gospodarstvo, politika
- vzgoja in izobraževanje
- narava, dom, hišni ljubljenci
- ljudje, družina, moški, ženske, otroci, mladina
- zdravje, hrana
- posel, finance
- prosti čas, glasba, film, razvedrilo, moda
- šport, turizem
- kultura, umetnost
- religija, duhovnost
- računalništvo, avtomobilizem itd.

Ko smo po metodi tematskega modeliranja Gigafido primerjali s prvo različico spletnega korpusa slovenščine slWaC (Logar Berginc in Ljubešić 2013), smo ugotovili, da imata korpusa izmed dvajsetih *skupnih* osem tem, *deloma skupnih* sedem tem, pet tem pa je bilo *različnih* (ibid.: 92):

> V Gigafidi so opaznejše teme naselje in cestni promet (zlasti z vidika prometnih nesreč), prireditve (zlasti z vidika njihove najave, opisa), televizijski in radijski program, neekipni športi ter zaposlitev. V slWaCu izstopajo film, glasba, potovanja in turizem, zunanja politika (zlasti EU, Hrvaška) ter mali oglasi.

Iz Tabel 1 in 2 v naslednjem poglavju (erjavec et al. 2015) je mogoče povzeti še podobnosti in razlike med temami Gigafide ter temami sodobnejše različice korpusa slWaC$_2$, nastalega leta 2014 (Erjavec in Ljubešić 2014):

a) Trinajst tem je skupnih: *človek, moški, ženska, družina, življenje*; *družba, RAZNO*; *šport*; *notranja politika*; *izobraževanje*; *finance*; *lokalna (prostorska) politika*; *pravo*; *publikacije, kultura, umetnost*; *avtomobilizem*; *zdravje*; *informacijsko-komunikacijska tehnologija* in *hrana*.

b) Tri teme so deloma skupne: *gospodarstvo* (Gigafida) – *gospodarstvo, razvoj* (slWaC$_2$); *prireditve v lokalnem prostoru* (Gigafida) – *prireditve (film, glasba, gledališče)* (slWaC$_2$); *živali, narava, bivalno okolje* (Gigafida) – *bivalno okolje* (slWaC$_2$).

c) Štiri teme so različne:

- Gigafida: *vojna, terorizem, kazniva dejanja*; *TV- in radijski program*; *promet*; *mediji*;
- slWaC$_2$: *potovanja, turizem*; *spletno nakupovanje*; *religija* in *svetovni splet*.

V primerjavi s spletnim korpusom slovenščine je torej v Gigafidi manj besedil o filmskih, glasbenih in sorodnih prireditvah, potovanjih in turizmu, malih oglasov, besedil o zunanji politiki, povezani z EU, spletnem nakupovanju in svetovnem spletu nasploh ter religiji. Z izjemo zadnje je pri vseh mogoče sklepati, da gre za teme, ki se v zadnjih letih v večjem obsegu objavljajo v spletnem mediju, kar govori v prid vključevanju spletnih besedil (s temi temami) v referenčni korpus. Analiza je potrdila tudi preveliko zastopanosti TV- in radijskih programov v Gigafidi (ki jo bo treba zmanjšati z deduplikacijo) in ustreznost seznama tem, ki smo ga pripravili pred zbiranjem, s tem da ga velja pri nadgradnji korpusa dopolniti še s temami pravo (zlasti v smislu javne uprave), promet, bivalno okolje in svetovni splet.

7 DODATNE TAKSONOMSKE KATEGORIJE

Gigafidina taksonomija je dokaj preprosta: besedila so na prvi ravni ločena na *tiskana* in *internetna*, v okviru tiska pa nato še na *knjižna* in *periodična*. Knjižna dela so po (ne)fikcijskosti vsebine razdeljena na *leposlovje* in *stvarna besedila*, periodično izdajana besedila pa na *časopise* in *revije*. Kategorija *drugo* je raznorodna (v Gigafido prinaša 0,67 % besed), združuje pa zapise sej Državnega zbora RS ter podnapise in postprodukcijska besedila RTV Slovenija. Prim.:

tisk
    knjižno
        leposlovje
        stvarna besedila
    periodično
        časopisi
        revije
    drugo
internet

Za splošno iskanje po korpusu se zdi, da taka taksonomija zadošča, za leksikografske potrebe pa bi bilo koristno, če bi jo dopolnili in/ali podrobneje razčlenili. Zgoraj smo v zvezi s tem že nakazali potrebo po ločeni kategoriji za učbenike in podobna besedila, v tem poglavju pa bomo v nadaljevanju na ta način razmišljali o spletnih besedilih v nestandardni slovenščini (blogovski zapisi, forumska sporočila, tviti, komentarji pod prispevki na novičarskih portalih). Dosedanje analize pa so pokazale tudi, da bi dodatna označenost korpusa leksikografom lahko pomagala pri odločanju o pripisu področnih in stilnih oznak.

7.1 Korpusni metapodatki in področne oznake v slovarju

Pripis področnih oznak tipa *sadjarstvo*, *avtomobilizem*, *bančništvo* ob slovarsko iztočnico oz. njen posamezen pomen je tesno povezan s terminologijo v splošnem slovarju. Če bi imela Gigafida vsaj del besedil označen s tematskimi kategorijami, bi te leksikografa lahko opozorile na morebitno področno poimenovalnost iztočnice, ki jo ureja, obenem pa bi taka označenost že v korpusu samem omogočala dodatna podkorpusna iskanja oz. bi še dodatno tematsko ali področno opredelila rezultate korpusnih poizvedb.[25] Kot smo ugotavljali že v Logar in Ljubešić (2013: 80), ima več tujih korpusov tematsko kategorijo pripisano pri strokovnih besedilih:

a) V Češkem nacionalnem korpusu SYN2010[26] so strokovna besedila členjena na:

- religijo
- pravo
- umetnost
- ekonomijo
- tehnologijo
- naravoslovje
- humanistiko in življenjske stile

b) V Hrvaškem nacionalnem korpusu[27] so načrtovali členitev:

- znanstvenih besedil na:
  - naravoslovne znanosti

25 Besedilna tema ali predmetno področje sta deloma prekrivna pojma (prim. za angleščino: *topic*, *domain*, *subject area*, *subject field*). Če se leksikologija bolj nagiba k področnim oznakam, je pri korpusih (pol)avtomatsko lažje določiti temo, ki pa je seveda lahko lastna več strokovnim področjem. V nadaljevanju bomo zato govorili o dodatni tematski označitvi korpusnih besedil kot eni od (novih) taksonomskih kategorij.

26 http://ucnk.ff.cuni.cz/english/syn2010.php (dostop 6. 7. 2015).

27 http://hnk.ffzg.hr/struktura.html (dostop 6. 7. 2015).

    - tehnične znanosti
    - biomedicinske znanosti
    - biotehnične znanosti
    - družboslovne znanosti
    - humanistične znanosti
- strokovnih besedil pa na:
    - potopise
    - kritike
    - medije
    - kriminalistiko
    - šport
    - politiko
    - ekologijo, bioetiko itd.

c) V Britanskem nacionalnem korpusu[28] pod informativno najdemo:

- svetovno politiko
- trgovino in finance
- umetnost
- religijo in filozofijo
- prosti čas itd.

Izhodiščna, čeprav ne v celoti uresničena tematska členitev je npr. značilna še za referenčni korpus Oxford English Corpus,[29] ki ga sestavlja dvajset delov, pretežno poimenovanih po temi, npr. računalništvo, okolje, prosti čas, vojska, transport. Ti deli so nadalje razdeljeni še na podteme oz. podpodročja (tako jih ima npr. šport kar okrog štirideset).

Pri dokončnem naboru tematskih kategorij, ki bi jih pripisali besedilom nadgrajene Gigafide, bi bilo smiselno imeti v razvidu tudi rezultate primerjav med Gigafido in slWaCom po metodi tematskega modeliranja (gl. razdelek 6.2 zgoraj in poglavje IV), pred pripravo tematske sheme pa za vsak korpusni dokument po metodi TF-IDF (angl. *Term Frequency – Inverse Document Frequency*; Salton in Buckley 1988) pridobiti še ključne besede. S pripravljeno tematsko shemo bi nato ročno označili učno množico dokumentov in izvedli strojno učenje ter nato korpus avtomatsko označili.

28 http://www.natcorp.ox.ac.uk/ (dostop 6. 7. 2015).
29 http://oxforddictionaries.com/words/the-oec-composition-and-structure (6. 7. 2015).

7.2 Korpusni metapodatki in stilne oznake v slovarju

Izpis stilnih oznak iz trenutne različice Leksikalne baze za slovenščino je pokazal, da so redaktorji pomene kvalificirali z naslednjimi pripisi v petih skupinah (Krek et al. 2013b: 94–96):

a) čas: *manj pogosta raba*, *beseda se v sodobni slovenščini v tem pomenu zelo redko uporablja*, *zastarelo*[30]

b) konotacija: *za izražanje poudarka*, *preneseno*, *odklonilno*, *izraža prizadetost*, *slabšalno*, *navadno z neodobravanjem*

c) kontekst: *v novinarskem žargonu*, *v oglasnih besedilih*, *pogosto v malih oglasih*, *zlasti v športu*, *v krščanstvu*, *v političnem kontekstu*

č) pragmatika: *kot pregovor*, *z neodobravanjem*, *evfemično*, *navadno kot zmerljivka*, *grobo in nekoliko prostaško*

d) register: *v zelo neformalnih situacijah*, *v neformalnih situacijah, v govoru*, *v neformalnem šolskem govoru*, *neformalno*

Za določitev konotacijskih in pragmatičnih oznak mora leksikograf ovrednotiti neposredno besedilno okolje, pri čemer mu lahko izdatno pomagajo orodja, kakršno je Sketch Engine[31] (Kilgarriff et al. 2004), neposredno pa si je s trenutnimi korpusnimi metapodatki mogoče delno pomagati pri časovno-frekvenčnih, kontekstualnih in registrskih oznakah.

a) Čas in frekvenca
Časovna neaktualnost besedišča z vidika sodobnosti se v korpusu neposredno iz metapodatkov (letnica izida) ne vidi, saj so v Gigafido vključena le besedila, izdana po letu 1990 (pretežno pa po letu 1996). To pomeni, da lahko leksikograf časovno oznako poda le na podlagi pregleda neposrednega besedilnega okolja v kombinaciji z analizo frekvenčnega razmerja med sopomenkami. Po drugi strani pa je Gigafida z besedili iz 20-letnega obdobja dovolj relevantna, da omogoča utemeljen prikaz porasta oz. upada pogostosti v rabi.[32] Pri zadnjem moramo biti ob osnovni frekvenci pozorni še na trend in na to, da moramo naraščanje ali upadanje pogostosti v nekem časovnem obdobju kombinirati še z razpršenostjo virov, relativno frekvenco glede na število besed/leto v korpusu in frekvenco morebitne sopomenke, katere porast ali upad v rabi je najbrž obraten. Težnja po prehajanju iz označevanja časovnosti v označevanje frekvenčnosti se pravzaprav vidi že pri preliminarnem naboru oznak v trenutni leksikalni bazi (prim. *manj pogosta raba*, *beseda se uporablja redko*).

30 Navajamo le po nekaj primerov iz preliminarne redakcijske faze (več gl. v poglavju VIII).
31 http://www.sketchengine.co.uk/ (dostop 6. 7. 2015).
32 Najbolj pregledno v obliki grafov, bolj strnjeno pa v obliki oznak.

b) Kontekst
Trenutne kontekstualne oznake so v leksikalni bazi raznorodne. Deloma so vezane na analizo neposrednega besedilnega okolja, pri čemer v manjši meri pomagajo tudi že obstoječi korpusni metapodatki (npr. leksikalne enote iz zapisov sej Državnega zbora RS). Dodatna označitev korpusa pri določanju takih oznak ne bi pomagala. Deloma pa so kontekstualne oznake povezane s temo (zgoraj: *zlasti v športu*, *v krščanstvu*, *v političnem kontekstu*), o čemer smo pisali že v razdelku 7.1.

c) Register
Registrske oznake enako kot kontekstualne deloma izhajajo iz analize neposrednega besedilnega okolja. Zdi se, da gre zlasti za prepoznavanje neformalnih govornih položajev, ki se lahko pojavljajo v vseh vrstah besedil: npr. v *leposlovju* v dialogih oseb, *v revijah* in *časopisih* v navedkih, intervjujih, polliterarnih žanrih ali literarnih podlistkih. Izhodiščna govorjenost je sicer značilna za dve vrsti besedil v Gigafidi (zapise sej Državnega zbora RS in podnapise RTV Slovenija), obe sta v taksonomiji označeni z *drugo* in poimenovani, kar lahko leksikografu neposredno pomaga pri določitvi registra. Tretji za registrske oznake zanimiv vir, ki je prav tako poimenovan, pa je *internet*, in sicer zlasti tista besedila, ki prihajajo z novičarskih portalov, natančneje: besedila komentarjev pod prispevki na novičarskih portalih. Novičarski portali, vključeni v trenutno Gigafido, so: 24ur.com, rtvslo.si, siol.net, arhivo.com, govori.se, najdi.si (novice), n-tv.si, pozareport.si, primorske.si in revija-reporter.si. Prvi trije portali so navedeni poimensko, preostali imajo skupno poimenovanje *internet – novice*. Pri dopolnitvi Gigafide z internetnimi besedili (gl. v nadaljevanju tega poglavja) bi bilo komentarjem pod prispevki na novičarskih portalih tudi zaradi lažjega leksikografskega prepoznavanja registrskih posebnosti koristno dodeliti ločeno taksonomsko kategorijo.

8 SKLEP

Pri gradnji Gigafide je sodelovalo 32 raziskovalcev z osmih znanstvenoraziskovalnih ustanov in ene založbe (Logar 2014: 4). Skoraj dve desetletji nastajajoči korpusi iz »serije FIDA« so primeri dobre prakse, ki so ažurno sledili evropskim korpusnojezikoslovnim dognanjem, zato je pri pripravi novega referenčnega korpusa slovenščine mogoče dobro začeti že kar tam, kjer smo z Gigafido končali, upoštevajoč spremembe, ki jih je v zadnjih letih v jezik in besedilno produkcijo prinesla nova digitalna družbena realnost, ter predloge izboljšav, ki so jih izpostavile ocene leta 2012 zaključenega izdelka. V prispevku se nismo opredeljevali do deležev posameznih vrst besedil v prihodnjem korpusu sodobne slovenščine,

za katerega je smiselno, da bi nastal iz Gigafide, prav tako npr. nismo predlagali seznamov besedil, ki manjkajo pri posameznih temah, nismo določali spletnih strani, na katerih bi izvedli pajkanje, ali pripravili nove taksonomije. Na ta in sorodna vprašanja mora odgovoriti konkretnejši dokument: specifikacija postopkov zbiranja besedil, ki pa ga je mogoče in smiselno pripraviti šele, ko imamo pred sabo konkreten projekt ter so znani njegovi časovni in finančni okviri.

Relevantnost gradiva glede na slovarski koncept je temeljnega pomena, smo zapisali v uvodu. Nobeden od obeh obstoječih konceptualnih predlogov novega slovarja ta hip še ni dokončen. Eden načrtuje izdelek »v smislu temeljnega in vsestranskega slovarskega priročnika za slovenščino v digitalni dobi«, ki bo »konceptualno in gradivno zasnovan povsem na novo« (Krek et al. 2013b: 20), drugi pa bo s slovarjem »nadalj/eval/ tradicijo *Slovarja slovenskega knjižnega jezika* v smislu aktualnosti jezikoslovne misli in opisa jezikovne rabe« (Gliha Komac et al. 31. 3. 2015: 1). Gigafida osnovnemu izhodišču – torej povsem novemu opisu sodobne slovenščine na podlagi splošne rabe – zadošča in ga omogoča, v skladu s tu ter v naslednjih poglavjih prikazanimi ugotovitvami in smernicami pa jo je mogoče – in treba – nadgraditi. Zadnje prilagoditve bo nato določil še dokončni slovarski koncept; od razvidnosti in doslednosti leksikografskega postopka pa je nato odvisno, kako bodo njeni podatki interpretirani ter v kolikšnem obsegu bodo upoštevani, izkoriščeni ali prezrti.

Nadgradnja Gigafide: spletna besedila

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, Vesna Mikolič

Abstract
This paper discusses the expansion of the Gigafida corpus, a Slovenian reference corpus, to include Internet content, i.e. webpages and user-generated content (tweets, blogs, forums and comments on news portals). The resources and tools available which are best suited to achieve this objective are discussed, and the Web crawling methodology used for this purpose is presented.

Keywords: reference corpus, Slovenian, dictionary, Internet content, Web crawling

Ključne besede: referenčni korpus, slovenščina, slovar, spletna besedila, pajkanje

1 UVOD

Leta 2012 smo poglavje z naslovom Spletna besedila v korpusu Gigafida (Logar Berginc et al. 2012: 45) začeli z ugotovitvijo, da postaja podajanje pisnega jezika v javni rabi vse manj domena tiska in vse bolj domena elektronskih medijev, ter ob tem zapisali podatek, da se je oktobra 2007 za internetne uporabnike v populaciji od 12 do 65 let v raziskavi RIS izreklo 66 % vprašanih. Novejši deleži so – pričakovano – še višji: po analizi Statističnega urada RS je imelo v prvem četrtletju leta 2014 dostop do interneta 97 % slovenskih gospodinjstev z otroki in 70 % gospodinjstev brez otrok, internet pa je v tem obdobju uporabljalo 72 % vseh oseb, starih od 16 do 74 let. Dodamo lahko še, da je

> 81 % teh oseb /.../ internet uporabljalo vsak dan ali skoraj vsak dan. Ti so ga v največjem odstotku (87 %) uporabljali za pošiljanje ali prejemanje e-pošte in za iskanje informacij o blagu ali storitvah. / 58 % oseb je v prvem četrtletju 2014 sodelovalo v družabnih spletnih omrežjih (v prvem četrtletju 2013 je bilo takih 53 %) (ibid.).

Pomemben podatek je tudi ta, da je 66 % uporabnikov do interneta dostopalo prek mobilnega telefona ali druge mobilne naprave (npr. bralnika). Internet je torej postal dostopen kjerkoli, in to ne le za branje, gledanje ter poslušanje, temveč tudi za pisanje in objavljanje besedil, slik, glasbe ipd. K široko dostopni javni besedi – nekoč omejeni na tisk, radio in televizijo – se tako lahko priglasi vsakdo, to pa v slovenščino v javni rabi prinaša nov segment: besedila, katerih jezikovna podoba kaže značilnosti, prej vezane predvsem na zasebne govorne položaje.

Oblikovalci aktualnih tujih referenčnih korpusov spletna besedila v korpuse vključujejo zelo različno. Pregled v Logar Berginc in Ljubešić (2013) je pokazal, da »skupna težnja, da bi se v referenčne korpuse vključevalo besedila z interneta in v kolikšnem obsegu bi to bilo, še ni jasno razvidna, če pa korpus že vsebuje ali bo vseboval besedila z interneta, se vanj v glavnem zajemajo besedila različnih žanrov« (ibid. 103). Tako imamo na eni strani npr. Oxford English Corpus, iz katerega nastajajo angleški slovarji Oxford, ki je skoraj v celoti sestavljen iz besedil s spleta, na drugi strani pa npr. Slovaški nacionalni korpus, na osnovi katerega se pravkar pripravlja Slovar sodobnega slovaškega jezika, ki spletnega dela sploh ne vsebuje (več gl. v Tabeli 1 v predhodnem poglavju).

Kot bo razvidno iz nadaljevanja poglavja, besedila s spletnih strani, komentarje pod prispevki na novičarskih portalih, blogovske zapise, tvite in forumska sporočila tu razumemo kot tvoren del javne pisne slovenščine, zaradi česar bi jih bilo treba zajeti tudi v korpus, ki bo podlaga za prihodnji referenčni slovar. Namreč: leksikografe bi morala zanimati leksika, ki jo v različnih okoliščinah

uporabljajo in ustvarjajo *vsi* govorci slovenščine, ne zgolj novinarji, prevajalci, pisatelji ipd. Tovrstno posebno pozornost zato danes terja prav (pol)javno pisno spletno komuniciranje, ki ga določajo okoliščine, kot so (ne)interaktivnost, (a)sinhronost, fizična (ne)prisotnost sogovornika in drugi situacijski dejavniki, katerih posledica je zelo interaktivna oblika komuniciranja z več prvinami spontanega govorjenega jezika ter z (za računalniško komuniciranje prilagojenimi) parajezikovnimi in prozodičnimi elementi (Crystal 2001). Naloga korpusa kot slovarskega vira torej mora biti zajem tudi te jezikovne realnosti, zato jo v poglavju osvetljujemo s štirih zornih kotov:

a) izhodiščnega stanja v Gigafidi (primerjalno s spletnim korpusom slovenščine slWaC$_2$),

b) raznovrstnosti spletnih besedilnih žanrov in njihove korpusne aktualnosti,

c) virov in orodij, ki so za tovrstno prihodnjo nadgradnjo Gigafide že na voljo (projekt JANES), ter

č) najustreznejše metodologije pajkanja, vključno s komentarjem možnosti gradnje spremljevalnega podkorpusa.

2 GIGAFIDA IN slWaC$_2$: STANJE, PRIMERJAVA, POVEZLJIVOST

Spletne strani, ki so bile vključene v Gigafido, in tehnologije za njihov zajem so natančneje opisane v že omenjenem poglavju knjige Logar Berginc et al. (2012: 45–67), zato naj spomnimo le, da je šlo pri vključevanju spletnih vsebin v Gigafido »v metodološkem smislu za prvi večji /dotedanji/ tak poskus pri nas, ki bi lahko oblikoval smernice za prihodnjo gradnjo referenčnih korpusov slovenščine ter nakazal nekatere zanimive (besedilnozvrstnoprimerjalne) jezikoslovne analize« (ibid.: 45). Gigafida tako vsebuje besedila 10 novičarskih portalov ter skupno 91 predstavitvenih spletnih strani podjetij (29) in ustanov (62). Pajkanje je bilo izvedeno v obdobju april 2010–april 2011, v korpus pa je prineslo več kot 185 milijonov besed, od kateri jih 63 % prihaja z novičarskih portalov (24ur.com, rtvslo.si, siol.net ipd.), 30 % s strani ustanov (gov.si, uni-lj.si, sazu.si, ijs.si itd.) ter 7 % s strani podjetij (eles.si, gorenje.si, kolosej.si itd.). Postopek za zajem besedil s spletnih strani je vseboval več korakov: izbor in pripravo programa za tri režime pajkanja (dnevno, mesečno ter enkratno), odstranjevanje spremnih in vnaprej pripravljenih besedil, detekcijo jezika ter na koncu še detekcijo dvojnikov in približnih dvojnikov. Izkazalo se je, da je za uresničitev na videz dokaj preproste naloge, tj. naloge vključitve spletnih besedil v referenčni korpus, potrebna dokaj kompleksna metodologija, ki pa smo jo – vključno z merili izbora spletnih strani in oceno pridobljenega – uspešno

preizkusili ter prilagodili slovenščini (več o najnovejših metodah pajkanja gl. v razdelku 5).

Prav v času vključevanja spletnih besedil v Gigafido – leta 2011 – pa je nastal še en, v tem segmentu metodološko podoben korpus slovenščine: korpus slWaC (Ljubešić in Erjavec 2011), ki je bil leta 2014 nadgrajen v slWaC$_2$ (Erjavec in Ljubešić 2014). Korpus slWaC$_2$ vsebuje 1,2 milijardi pojavnic iz besedil, pridobljenih z dobrih 37.000 spletnih domen oz. 2,8 milijonov naslovov URL. Metodologija gradnje obeh različic korpusa slWaC je podrobneje predstavljena v obeh navedenih virih ter v Logar Berginc in Ljubešić (2013: 87–89).

Obstoj dveh obsežnih korpusov slovenščine je že spodbudil nekatere primerjave, ki so pokazale, kaj v enem in drugem korpusu *je*, omejeno (kolikor jih pač daje primerjava dveh entitet) pa tudi, kaj v obeh korpusih manjka. Primerjava po metodi frekvenčnega profila (Rayson in Garside 2000) med Gigafido ter korpusom slWaC$_2$ (Erjavec et al. 2015: 40) je tako med drugim pokazala, da je v slednjem več besedil, povezanih z računalništvom in spletom, ter uporabniških spletnih vsebin, medtem ko Gigafida vsebuje več besedil o športu, predvsem pa več za časopise značilnih besedil o notranji politiki, gospodarstvu in kaznivih dejanjih. Primerjava je pokazala tudi tehnične razlike, saj je bila Gigafida označena s programom Obeliks (Grčar, Krek in Dobrovoljc 2012), medtem ko smo za slWaC$_2$ uporabili program ToTaLe (Erjavec et al. 2005). Ker se orodji mestoma razlikujeta v označevanju, so se nekatere besede izpostavile kot zelo ključne za slWaC$_2$, ne pa za Gigafido, čeprav gre v resnici samo za razliko v obdelavi obeh korpusov. Tako je *le-ta* enkrat obravnavan kot ena pojavnica, drugič kot tri, *več* pa je npr. enkrat lematiziran kot »veliko«, drugič pa kar kot »več«. Pri kakršnemkoli povečevanju Gigafide bo zato treba pri izbiri orodij za jezikovno označevanje imeti v mislih smiselnost uporabe enotnega orodja za vse (prihodnje) slovenske korpuse (Erjavec et al. 2015).

Že objavljenim primerjalnim podatkom (Erjavec et al. 2015 ter Logar Berginc in Ljubešić 2013) tokrat dodajamo še podatke iz novejše primerjave med Gigafido in korpusom slWaC$_2$, narejene po metodi tematskega modeliranja (Blei et al. 2003; Sharoff 2010); s tem da bomo tu pri tematskih profilih obeh korpusov pozorni le na morebitne šibkosti Gigafide (Logar 2015). Tabeli 1 in 2 tako kažeta 20 za Gigafido in slWaC$_2$ najbolj značilnih tem.

Tabela 1: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po teži v korpusu Gigafida.

Tema	Teža*	Samostalniška lema
človek, moški, ženska, družina, življenje	4,835	otrok leto dan čas ženska življenje človek družina oče moški roka prijatelj glava žena mama mož sin starš hiša
šport	4,034	tekma mesto leto ekipa zmaga točka igra sezona igralec prvenstvo klub liga prvak trener minuta konec pokal krog reprezentanca
notranja politika	3,639	predsednik vlada država stranka svet minister leto zakon volitev predlog poslanec vprašanje komisija član odbor zbor seja politika ministrstvo
družba, RAZNO	3,631	človek življenje svet čas odnos način stvar država vprašanje družba primer beseda delo moč stran problem resnica leto občutek
prireditve v lokalnem prostoru	2,865	ura društvo leto prireditev dan sobota dom član vas mesto občina skupina nedelja šola srečanje gost obiskovalec dvorana delo
vojna, terorizem, kazniva dejanja	2,669	leto vojna država policija policist človek vojska dejanje orožje dan napad vojak žrtev sodišče oblast zapor kazen čas mesto
TV- in radijski program	2,622	film leto glasba oddaja tv poročilo skupina serija dan pesem festival čas koncert predstava program vloga gledališče del novica
promet	2,481	cesta pot dan nesreča leto ura voda voznik morje vozilo mesto meter promet letalo kilometer čas vožnja kraj avtomobil
gospodarstvo	2,47	leto odstotek država podjetje cena trg plača izdelek rast razvoj delo gospodarstvo proizvodnja delavec število področje strošek mesec sistem
izobraževanje	2,363	šola delo leto otrok program področje znanje študent projekt izobraževanje univerza fakulteta učenec razvoj starš organizacija učitelj center zavod
finance	2,292	milijon evro tolar leto banka družba podjetje odstotek delnica milijarda dolar denar vrednost cena prodaja delež dobiček trg sklad
lokalna (prostorska) politika	2,228	občina leto prostor gradnja objekt cesta projekt območje delo zemljišče mesto milijon stanovanje okolje podjetje načrt denar voda tolar
živali, narava, bivalno okolje	2,153	žival barva prostor vrsta pes voda hiša gozd del material les vrt tla drevo konj čas leto vrata oblika

Tema	Teža*	Samostalniška lema
pravo	2,145	zakon člen sodišče postopek pravica primer podatek organ odstavek dan oseba podlaga pogodba delo odločba sklad stranka določba zadeva
publikacije, kultura, umetnost	2,087	leto knjiga delo razstava stoletje cerkev mesto čas muzej svet ime zbirka umetnost avtor jezik zgodovina del slika beseda
avtomobilizem	2,021	m sit avtomobil motor km cena vozilo eur d e l model leto avto x n g r h
zdravje	1,942	bolezen zdravnik bolnik zdravilo telo človek zdravljenje leto koža težava dan zdravje primer rak bolnišnica bolečina kri celica čas
mediji	1,788	naslov stran številka medij novinar revija dan nagrada pošta časopis leto ime delo informacija oddaja članek televizija bralec vprašanje
informacijsko-komunikacijska tehnologija	1,491	računalnik sistem uporabnik podatek program slika stran naprava uporaba kartica telefon zaslon internet omrežje model oprema tehnologija možnost storitev
hrana	1,437	vino voda olje rastlina minuta meso sladkor g sol hrana zelenjava jed žlica okus sadje mleko krompir sok list

* »Teža« v drugem stolpcu pomeni razpršenost posamezne teme v korpusu.

Tabela 2: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po teži v korpusu slWaC$_2$.

Tema	Teža	Samostalniška lema
človek, moški, ženska, družina, življenje	3,929	otrok dan čas leto človek ženska roka pes življenje stvar prijatelj moški glava mama ura družina starš svet konec
družba, RAZNO	3,266	človek življenje svet čas način odnos stvar družba otrok ljubezen beseda primer vprašanje resnica pot občutek ženska moč problem
notranja politika	2,626	vlada država predsednik stranka zakon svet leto predlog minister član poslanec komisija vprašanje zbor odbor politika skupina pravica mnenje
potovanja, turizem	2,524	pot mesto dan cesta ura leto čas vrh morje voda smer gora meter del dolina kraj gozd hotel stran
gospodarstvo, razvoj	2,36	podjetje področje razvoj sistem projekt delo leto trg storitev okolje država cilj organizacija program izdelek znanje rešitev sodelovanje tehnologija
finance	2,265	leto evro odstotek milijon podjetje banka država cena družba denar trg vrednost rast milijarda sredstvo delnica plača prodaja mesec

Tema	Teža	Samostalniška lema
šport	2,232	tekma ekipa mesto igra leto točka zmaga sezona igralec minuta prvenstvo klub liga konec tekmovanje prvak rezultat trener pokal
prireditve (film, glasba, gledališče)	2,139	film leto glasba skupina album pesem festival koncert skladba čas oder nastop predstava nagrada vloga dan zasedba oddaja zgodba
izobraževanje	2,072	šola otrok leto delo program študent učenec znanje starš ura izobraževanje fakulteta univerza čas študij delavnica področje učitelj dan
zdravje	2,059	telo bolezen koža zdravilo težava zdravljenje zdravnik dan leto bolnik bolečina človek zdravje celica primer čas kri otrok učinek
spletno nakupovanje	2,042	stran podatek uporabnik naslov storitev vsebina račun pošta cena ime nakup internet številka informacija izdelek naročilo ponudba dan paket
pravo	2,016	člen zakon sodišče postopek pravica odstavek oseba pogodba primer dan stranka podlaga sklad organ delo določba odločba podatek pogoj
lokalna (prostorska) politika	1,937	občina leto projekt društvo območje mesto delo prostor sredstvo objekt program član center gradnja zavod organizacija področje okolje ministrstvo
religija	1,887	leto cerkev človek vojna bog življenje dan mesto čas smrt vojska svet država oče maša ime beseda vera stoletje
publikacije, kultura, umetnost	1,826	leto knjiga delo jezik razstava avtor medij beseda fotografija nagrada zbirka revija del umetnost zgodba naslov čas svet dogodek
informacijsko-komunikacijska tehnologija	1,781	računalnik naprava sistem slika program telefon fotografija podatek uporabnik uporaba video zaslon stran aplikacija dokument model kamera različica oprema
avtomobilizem	1,697	vozilo avtomobil motor barva vožnja voznik model kolo avto del leto oblačilo znamka cesta hitrost obleka sedež oprema sistem
bivalno okolje	1,624	voda prostor energija hiša material sistem površina odpadek zrak objekt naprava del uporaba stanovanje temperatura okno okolje les plin
hrana	1,471	hrana voda olje rastlina vino mleko okus meso zelenjava vrsta jed sadje izdelek oseba količina dan kislina žival sladkor
svetovni splet	0,52	piškotek dan nastavitev seja mesto namen stran storitev uporaba informacija podatek oglaševanje klik gumb primer ura facebook možnost novica

V tabelah lahko prepoznamo tri teme, ki jih bo treba še posebej imeti v mislih pri izboru naslovov URL, s katerih bodo pri nadgradnji Gigafide pridobivana nova spletna besedila (o tvitih, forumskih sporočilih, komentarjih pod prispevki na novičarskih portalih in blogovskih zapisih gl. v nadaljevanju). Gre za teme: *potovanja, turizem*; *spletno nakupovanje* in nasploh *svetovni splet* (gl. zadnjo vrstico v Tabeli 2). Tema *religija* je edina, ki bi se jo dalo bolje vključiti v Gigafido že tudi s tiskanimi deli (je pa pri tem ključen odziv besedilodajalcev).

Samo po sebi se ob koncu primerjav v razmislek ponuja vprašanje morebitne kar neposredne vključitve spletnega korpusa slovenščine slWaC$_2$ v novo Gigafido. Z vidika bolj usmerjenega in kontroliranega, pa tudi časovno predvidljivega ter enakomernega zajema besedil z izrecnim namenom vključitve v referenčni korpus je na to vprašanje bolje odgovoriti negativno, ni pa treba, da bi bili tudi prihodnji nadgradnji obeh korpusov povsem ločeni. Nasprotno: kot je razvidno v razdelku 5, ju tesno povezuje metodologija izdelave, obstoj dveh korpusov sodobne slovenščine pa je koristen tudi v smislu medsebojnega dopolnjevanja ter izkaza medsebojnih razlik in pomanjkljivosti.

3 SPLETNA BESEDILNA ŽANRSKOST IN SLOVARSKI VIRI

Na spletu kot najvplivnejšem mediju 21. stoletja se srečujemo z različnimi komunikacijskimi okolji oz. področji in vsemi štirimi osnovnimi funkcijami besedil (Skubic 1995; Mikolič 2007): spoznavno, sporazumevalno, izvršilno ter umetnostnoizrazno. Na spletu se oblikujejo različne diskurzivne/govorne skupnosti, ki se v okviru določenega diskurza/govora odločajo za zanj značilne jezikovne izbire. Nekatere funkcije spletnih besedil so tako vezane na (bolj) neformalne govorne položaje in se udejanjajo v besedilih s številnimi nestandardnimi jezikovnimi prvinami, druga spletna besedila pa ustrezajo pojmu javnega komuniciranja v ožjem pomenu besede (Škiljan 1999) in sooblikujejo jezikovni standard. Jezikovna heterogenost spleta seveda močno vpliva na spremembe v jeziku in širjenje njegove leksike, zato je nujno ugotoviti, katera besedila morajo biti sestavni del korpusa, ki bo temeljni vir za sodobni slovar slovenskega jezika (ter hkrati, katerim se je mogoče (zaenkrat) odreči).

Opis spletne zvrstnosti in njenih ključnih dejavnikov je pravzaprav zelo nehvaležna naloga, tako zaradi obsežnosti oz. neobvladljivosti gradiva kot zaradi maloštevilnih raziskav spletnih žanrov in njihovih ciljnih javnosti (Crowston 2010: 17, 26). Kljub temu lahko na osnovi pregledane literature (Bishop 2009; Crowston 2010; Domingo in Heinonen 2008; Herring et al. 2004; Oblak et al. 2005) ugotovimo, da pri vseh avtorjih izstopata predvsem dve ključni merili, pomembni

tako za analizo spletne besedilne zvrstnosti kot za utemeljitev izbora spletnih besedil za korpus:

- avtorstvo oz. razmerje med sporočevalcem in naslovnikom (eden ali več avtorjev, isti, znan vir informacij ali več različnih, tudi anonimnih virov, formalno ali neformalno razmerje, ki zahteva, da je diskurz ne glede na število avtorjev in virov informacij bolj ali manj formalen in notranje konsistenten),
- funkcija ter z njo povezana notranja in zunanja zgradba oz. oblika besedila in večja ali manjša formalnost govora (tudi večkodnost, posodabljanje).

Z vidika avtorstva je osnovna delitev spletnih besedil na:

- klasične spletne strani (HTML), pri katerih je avtor en sam (npr. lastnik osebne spletne strani) ali je vir besedil isti in znan (npr. podjetje s svojo spletno stranjo) in pri katerih gre večinoma za enosmerno komuniciranje,
- žanre spletnih skupnosti (angl. *web-based community genres*; Bishop 2009), pri katerih besedila soustvarja več avtorjev in gre torej vedno za večsmerno komuniciranje,
- blogovske zapise kot vmesni žanr med eno- in dvosmerno obliko komuniciranja.

Za **klasične spletne strani** je v glavnem značilno enosmerno komuniciranje (najpogostejša izjema so tu sicer medijska spletna mesta, ki lahko vključujejo forumska sporočila, komentarje ali blogovske zapise bralcev). Vir besedil je pri spletnih straneh znan oz. lahko določljiv. Razmerje med sporočevalcem in naslovnikom je večinoma formalno, besedila nagovarjajo širšo javnost, zato je tudi diskurz v glavnem formalen. Med klasične spletne strani uvrščamo:

- spletne portale (tudi tipa Wikipedija, Wikivir, Wikiverza, Wikiknjige ipd.),
- medijska spletna mesta,
- komercialne in korporativne spletne strani,
- spletne strani vladnih in nevladnih organizacij ter lokalne samouprave.

Ena od oblik spletnih strani so tudi osebne spletne strani, ki pa so manj formalne in se lahko že približujejo obliki ter namenu blogovskega zapisa in žanrov spletnih skupnosti (npr. stranem na Facebooku).

Žanri spletnih skupnosti so h kolektivnemu delovanju usmerjena spletna mesta oz. interaktivne besedilne vrste računalniško posredovanega komuniciranja, pri katerem sodeluje več avtorjev, določajo pa ga prevladujoči akterji, komunikacijsko

okolje ali tema in notranja zgradba. Tudi jezikovne izbire tu določa narava interakcije med akterji, ti pa so zelo raznovrstni, pogosto tudi anonimni, zato je izraznost besedil zelo pestra, večinoma pa vključuje neformalne jezikovne prvine. Ker ti žanri vse bolj nadomeščajo govorno komuniciranje, gre pogosto za zapisan govorjeni jezik, pri nekaterih aplikacijah tudi za govorjena besedila. Med žanre spletnih skupnosti lahko uvrščamo besedila različnih spletnih orodij in družbenih omrežij, kot so:

- forumska sporočila (uporabniki razpravljajo na določeno temo),
- Twitter, Facebook, Myspace, LinkedIn (besedila, kot so: tviti, stanja oz. misli, komentarji stanj, fotografije, videoposnetki, hiperpovezave, skupine glede na interese, ustvarjanje dogodkov, povabila idr.),
- Instagram (objavljanje fotografij, povezave na Twitter ali Facebook),
- Ask.fm (uporabniki ustvarijo račun, ostali uporabniki jim postavljajo vprašanja, povezava na Twitter ali Facebook),
- Snapchat (mobilna aplikacija, prek katere uporabniki s svojimi prijatelji delijo misli, videoposnetke, fotografije itd., ki izginejo čez nekaj minut),
- Viber (mobilna aplikacija pametnega telefona, po kateri se komuniciranje odvija prek spleta, lahko je pisno ali govorno, vključuje pa imenik uporabnikov),
- spletne klepetalnice (različne kategorije – »sobe«, kjer se povezujejo uporabniki z enakimi interesi, namenjene so spoznavanju novih ljudi),
- komentarji novinarskih prispevkov, videoposnetkov itd. (razvijejo se v debato na določeno temo med ponavadi nepoznanimi uporabniki in delujejo na način foruma).

Blogovski zapisi so največkrat del publicistične besedilne vrste, namenjene širši javnosti in so pogosto v neposredni interakciji z njo. Avtor je ponavadi en sam in znan, saj se bralci največkrat prav zaradi avtorja odločajo za obisk bloga, kar veča njegovo branost in vpliv. Ker jih lahko pišejo profesionalni ali polprofesionalni pisci (novinarji in druge znane osebe iz javnega življenja, ki so bolj ali manj vešče javnega komuniciranja), pa tudi neprofesionalni avtorji, so jezikovne izbire odvisne od piščeve sporazumevalne zmožnosti, predvsem pa od vrste občinstva, ki ga želi pisec doseči. Po Domingo in Heinonen (2008) lahko ločimo naslednje vrste publicističnih blogovskih zapisov, ki se razlikujejo po profesionalnosti piscev in institucionaliziranosti okolja:

- državljanski blogovski zapisi (pišejo jih neprofesionalni pisci izven medijev),

- blogovski zapisi občinstev (pišejo jih neprofesionalni pisci v okviru medijev),
- novinarski blogovski zapisi (pišejo jih novinarji izven medijev),
- medijski blogovski zapisi (pišejo jih novinarji v okviru medijev).

Novinarskim blogovskim zapisom izven medijev so podobni blogi različnih oseb iz javnega življenja, ki so profesionalni ali polprofesionalni pisci (pisatelji, igralci, pevci, politiki ipd.). Enako lahko tudi besedila v četrti skupini, t. i. medijske blogovske zapise, pišejo tudi drugi profesionalni pisci, ne zgolj novinarji, npr. pisatelji, igralci, režiserji ipd.; tovrstni zapisi so pravzaprav redne rubrike z mnenji in komentarji (dokaj) stalnih avtorjev, ki ne izražajo nujno stališč medijev, v katerih gostujejo ali so pri njih zaposleni.

Z vidika *funkcije* besedila razvrščamo v skupine širših besedilnih zvrsti in ožjih besedilnih vrst, in sicer glede na naslednje skupne lastnosti: namen oz. vplivanjsko vlogo, naslovnika, referenco, zunanjo in notranjo zgradbo besedila ter z njimi povezano večjo ali manjšo notranjo konsistentnost in formalnost diskurza (prim. Mikolič 2013; Nidorfer Šiškovič 2013). Po teh lastnostih smo naredili preliminarno klasifikacijo besedil v spletnem okolju, pri čemer smo se naslonili na Crowstona (2010), ki povzema ključne tipologije spletnih besedil glede na namen in obliko. Tako lahko poskusimo besedilno zvrstnost na spletu opisati v okviru spodnjih skupin besedilnih vrst (povzeto po Mikolič in Rolih 2015), ki vse, razen pogovarjalnih z večjim obsegom vsebinskih in jezikovnih elementov zasebnega komuniciranja, svoj namen dosegajo v javnem komuniciranju s ciljnimi javnostmi ali pa širšo javnostjo, zato se v večini primerov v njih oblikuje formalni govor in standardni jezik:

- *Pogovarjalne in vsaj v delu zasebne besedilne vrste*: elektronska pošta, spletne klepetalnice, tviti in druga besedila družbenih omrežij (npr. Twitter, Facebook) ter forumska sporočila. Pripis »zasebnosti« pri teh besedilih s povezovalno-izrazno komunikacijsko vplivanjsko vlogo sloni na zanje značilnem večjem obsegu prvin zasebnega jezika oz. vsebinskih elementov zasebnih komunikacijskih sfer (Škiljan 1999), sociolektov in idiolektov (Skubic 2004).
- *Predstavitvene in promocijske besedilne vrste*: osebne spletne strani, spletne strani podjetij, ustanov, organizacij, društev ipd. Gre za besedila, namenjena širši javnosti, pri katerih se prepletata predstavitvena in usmerjevalna vplivanjska vloga. Osebne spletne strani imajo poleg predstavitvenega tudi samopromocijski namen, istočasno pa nastajajo z namenom vzpostavljanja družbenih mrež, saj želi avtor obiskovalcem svoje spletne strani sporočiti svoja stališča, poglede, poročati o svojem delu in s tem

vzpostaviti tesnejše medsebojne vezi. Predstavitveni namen spletnih strani različnih podjetij in organizacij se pogosto tesno prepleta s komercialnimi nameni.

- *Oglasne in komercialne besedilne vrste*: oglasna sporočila, zbirke povezav (angl. *link collections*), spletne trgovine, spletni portali za trženje in prodajo. Usmerjevalna vplivanjska vloga teh besedil se kaže v njihovem vplivanju na nakupno ravnanje naslovnikov.
- *Poročevalske in širšepublicistične besedilne vrste*: novinarska besedila različnih žanrov, spletne izdaje tiskanih medijev, prispevki, povezani z življenjskimi stili (kuharski recepti, nasveti v obliki t. i. tutorialov, vodiči za zdravo telo itd.).
- *Programerske besedilne vrste*: tehnični podatki/pomoč/podpora, poročila o težavah (angl. *problem reports*), pogosta vprašanja ali FAQ (angl. *frequently asked questions*). Tu gre za (poljudno)strokovna besedila, namenjena širši javnosti, ki jih na spletno stran postavijo upravljalci oz. programerji spletnih strani.
- *Akademske besedilne vrste* (dostopne npr. prek Googlovega Učenjaka): znanstvena in strokovna besedila s spoznavno in predstavitveno vplivanjsko vlogo, namenjena akademski ter strokovni javnosti.
- *Uradne in uradovalne besedilne vrste*: zapisniki sej državnih organov, zakonodajne strani, strani borze, pravilniki itd.; e-uprava, e-prijave ipd. Namen teh besedil z izvršilno vplivanjsko vlogo je seznaniti širšo javnost s ključnimi postopki, pravili in zakoni v državi ter hkrati omogočiti uradovanje z upravnimi organi prek spletnih obrazcev.
- *Literarne in polliterarne besedilne vrste*: umetnostna besedila, za katera je značilna skladnost z jezikovnim standardom z namernim odstopanjem od njega in individualizacijo jezikovnega stila. Najpogostejši polliterarni spletni vrsti sta blogovski zapis in spletni dnevnik.

Nedvomno je večina naštetih spletnih besedilnih vrst – čeprav še premalo raziskanih tako v slovenskem kot mednarodnem merilu – tudi v slovenščini zelo živih v smislu njihovega uresničevanja, branosti in vpliva tako v okviru različnih vrst družbenega komuniciranja kot v okviru razvoja jezika. Zaradi hitrega spreminjanja spletnih orodij lahko nekatere vrste hitro zastarajo (ta hip so v zatonu npr. spletne klepetalnice), namesto njih pa se pojavijo nove, lahko s podobnim ali povsem drugačnim namenom ter samosvojo jezikovno podobo. Prav zato je treba pri pripravi slovarskih opisov ne le *upoštevati* spletno jezikovno stvarnost, temveč ji tudi sproti *slediti*.

Ob predpostavki, da se bo iz nadaljnjih analiz spletnega gradiva zgornji nabor spletnih besedilnih vrst potrdil kot relevanten tudi za slovenski jezik, se zdi pri nadgradnji Gigafide utemeljeno upoštevati tisti del besedilnih vrst, ki imajo znanega avtorja oz. vir, svoj namen dosegajo v javnem komuniciranju in zato bolj neposredno sooblikujejo jezikovni standard. Kot taka se kažejo predvsem besedila vseh skupin klasičnih spletnih strani in tistih osebnih spletnih strani, ki so široko brane, dalje blogovski zapisi profesionalnih in polprofesionalnih piscev, tviti in Facebookove strani oseb ter ustanov, ki imajo večji vpliv na splošno jezikovno rabo (število sledilcev, medijski odzivi). Z vidika funkcije gre za del pogovarjalnih besedilnih vrst ter večji del predstavitvenih in promocijskih besedil, dalje poročevalske in širšepublicistične besedilne vrste, uradne in uradovalne ter literarne in polliterarne besedilne vrste – vse, kot že rečeno, ob pogoju večje vplivnosti in široke branosti ter na način, da bo iz taksonomskih kategorij korpusa vidno, za katere žanre gre.

4 NESTANDARDNO ZAPISANA SPLETNA BESEDILA PROJEKTA JANES

Kot smo zapisali že zgoraj, je za jezik pisnega spletnega komuniciranja značilna pogosta raba nestandardnih jezikovnih oblik, kot so nestandardni zapis besed in specifične krajšave. Zaradi tega je jezikoslovna analiza in posledično tudi avtomatska obdelava tovrstnih vsebin otežena (Sproat et al. 2001), prizadevanja za premostitev teh ovir pa so trenutno ena bolj vročih tem na področju računalniškega jezikoslovja. Leta 2014 se je tudi za sodobno slovenščino začelo delo na tem področju, in sicer v okviru temeljnega raziskovalnega projekta JANES (Jezikoslovna analiza nestandardne slovenščine), ki ima za enega od ciljev zgraditi vire in razviti orodja ter vzpostaviti metodologijo za proučevanje jezika spletnega komuniciranja (Fišer et al. 2014).

4.1 Korpus spletne slovenščine JANES

V trenutno različico korpusa JANES smo vključili štiri zvrsti uporabniških spletnih vsebin, in sicer tvite, forumska sporočila, komentarje pod prispevki na novičarskih portalih in blogovske zapise. Tviti se zajemajo z namenskim orodjem TweetCat (Ljubešić et al. 2014), pri čemer zajem poteka sproti, trenutno že več kot dve leti. Za zajem forumskih sporočil in novičarskih portalov, s katerih smo pridobili komentarje uporabnikov, smo izbrali po nekaj virov, ki so v slovenskem spletnem prostoru najbolj priljubljeni, ponujajo največ jezikovne produkcije in/ali predstavljajo pomemben del slovenskega spletnega prostora. Ker se spletna mesta po sestavi med seboj razlikujejo, smo uporabili ciljno pajkanje (več gl. v

razdelku 5.2) in za vsak vir posebej napisali ekstraktor besedila. Za gradnjo podkorpusa blogovskih zapisov smo uporabili kar deduplicirano različico splošnega korpusa slovenskega spleta slWaC$_2$, iz katerega smo zajeli vsa besedila, pri katerih se v domeni pojavi niz »blog«. Rešitev je začasna, saj za razliko od podkorpusov forumov in komentarjev zanje zaenkrat še nismo izdelali ciljnega ekstraktorja, tako da nimamo ohranjene notranje strukture blogovskih zapisov, npr. razdelitve na glavno besedilo in komentarje nanj.

Vsi našteti podkorpusi so bili združeni v korpus JANES, v katerem so poenoteni in s tem tudi poenostavljeni metapodatki posameznih besedil. Podkorpusi in korpus JANES so zapisani v formatu XML, ki omogoča strukturiranje korpusa, zapis metapodatkov in konsistenten zapis znakov po standardu Unicode. Korpus smo tudi jezikoslovno označili. Prvi korak označevanja sta bili tokenizacija in stavčna segmentacija, za kar smo uporabili standardno knjižnico mlToken za slovenski jezik, ki je del programa ToTaLe (Erjavec et al. 2005). V naslednjem koraku smo besedne pojavnice normalizirali z metodo, ki temelji na statističnem strojnem prevajanju črk, naučena pa je bila na 1.000 ključnih besedah iz korpusa tvitov glede na korpus Kres (Logar Berginc et al. 2012: 77–97) in na njihovih ročno normaliziranih oblikah (Ljubešić et al. 2014). Z orodji za standardno slovenščino programa ToTaLe smo nato normalizirane besede še oblikoskladenjsko označili in lematizirali (več o jezikovnem označevanju gl. v Erjavec et al. 2015a).

Korpus JANES, ki je, kot je razvidno iz zgornjega, omejen na javne spletne uporabniške vsebine, trenutno vsebuje dobrih 161 milijonov pojavnic. Največji delež v korpusu predstavljajo tviti (38 %), sledijo jim forumska sporočila (29 %), blogovski zapisi (24 %) in komentarji pod prispevki na novičarskih portalih (9 %). Že trenutni korpus je tako uporaben za leksikografske namene, saj je vsebinsko komplementaren z Gigafido, je zadosti velik in tudi razmeroma raznovrsten. Seveda pa bi bilo koristno še povečati število in raznovrstnost virov, predvsem pri forumskih sporočilih in komentarjih pod prispevki na novičarskih portalih.

4.2 Nestandardni jezik v korpusu JANES

Korpus JANES sicer vsebuje uporabniške spletne vsebine, nikakor pa niso vsa besedila v njem napisana v nestandardni slovenščini. Ročni pregled manjšega števila naključno izbranih tvitov je celo pokazal (Ljubešić et al. 2015), da je takšnih besedil v korpusu manj kot tretjina, česar sprva nismo pričakovali. A izkazalo se je, da družbena omrežja za obveščanje javnosti in promocijo uporabljajo tudi številna podjetja ter ustanove, kot so časopisne in druge medijske hiše, javne ustanove itd., ki tipično generirajo več besedil kot zasebni uporabniki ter v svojem uradnem komuniciranju uporabljajo standardno slovenščino.

Za proučevanje nestandardne slovenščine bi bila zato koristna ločitev spletnih (in drugih) besedil, ki so zapisana v nestandardnem jeziku, od vseh ostalih. Za ta namen smo razvili metodo za merjenje stopnje nestandardnosti (Ljubešić et al. 2015), pri kateri smo v prvem koraku analizirali večje število besedil in ugotovili, da je treba ločiti tehnično od jezikoslovne nestandardnosti. Prva pomeni npr. to, da pisec piše vse besede z malo začetnico ali ne uporablja ločil, druga pa npr. to, da besede zapisuje pogovorno, uporablja sleng ali drugo nestandardno leksiko. Za ti dve razsežnosti smo oblikovali navodila za označevalce, ki so nato ročno označili manjše vzorce iz podkorpusov JANES z oceno njegove tehnične oz. jezikoslovne nestandardnosti od 1 (zelo standardno) do 3 (zelo nestandardno).

Definirali smo okoli trideset značilk, za katere smo predpostavili, da bi lahko služile kot delne mere ene ali druge vrste nestandardnosti. Značilke so bodisi na nivoju znakov (npr. razmerje števila ločil glede na število vseh znakov v besedilu), na nivoju nizov (npr. razmerje besed z veliko začetnico do vseh besed) ali pa na nivoju leksike, pri čemer smo pozorni predvsem na razmerje leksike opazovanega besedila do besed v oblikoskladenjskem leksikonu Sloleks (Dobrovoljc et al. 2015), npr. razmerje kratkih besed, ki so v besedilu, a jih ni v Sloleksu. S temi značilkami smo na učni množici naučili regresor, ki lahko novim besedilom pripisuje obe meri nestandardnosti. S tem ko smo vsa besedila v korpusu JANES opremili z metapodatkom o stopnji tako tehnične kot jezikoslovne nestandardnosti, smo dobili možnost, da se pri korpusnih študijah osredotočimo samo na tisti del korpusa, ki je napisan v želenem tipu in stopnji (ne)standardnosti, leksikografu pa ta podatek lahko služi kot dobra redakcijska orientacija, saj lahko npr. za proučevanje nestandardnih zapisov besed iz korpusa izdvoji samo besedila, označena kot zelo jezikovno nestandardna.

5 METODOLOGIJA PAJKANJA

Pajkanje je proces, pri katerem z avtomatskimi metodami s spleta zajemamo dokumente bodisi za izdelavo indeksov spletnih iskalnikov, pridobivanje drugih informacij s spleta ali pa za izdelavo korpusov. Če je pri prvem najpomembnejši visok priklic, smo pri drugem bolj osredotočeni na pridobivanje jezikovnih vsebin, saj je bolje izgubiti dele zajetih dokumentov kot pa izdelati zelo šumen korpus, ki bi poleg zveznega besedila vseboval tudi elemente, kot so glave in noge spletnih strani, navigacijo itd.

Obstajata dva osnovna pristopa k pajkanju jezikoslovno zanimivih podatkov. Prvi, *generični* pristop, za vse zajete dokumente uporablja isti postopek obdelave in ima to prednost, da je enostavnejši za implementacijo ter pokriva zelo raznovrstne tipe dokumentov, ima pa tudi vrsto slabosti. Zajeti podatki so bolj šumni

in slabše strukturirani, saj npr. ne razlikujejo naslovov in podnaslovov od samega besedila, ravno tako pa ne vsebujejo potencialno koristnih metapodatkov o besedilu, kot so npr. datum in čas objave, ime avtorja ter razvrstitve dokumentov v vsebinske kategorije. Drugi pristop je *ciljni*. Pri njem je implementacija pajkanja prilagojena posameznemu izvoru dokumentov. Prednosti ciljnega pristopa so manjša količina šuma, boljša struktura zajetih besedil in večja količina metapodatkov, slabost pa je ta, da je treba program za zajem prilagoditi vsakemu izvoru posebej, kar je večinoma časovno zahteven proces.

Generični pristop se uporablja pri izdelavi velikih zbirk besedil, ki temeljijo bodisi na neki vrhnji spletni domeni (kot npr. ».si«) oz. posameznem jeziku (dober primer tega pristopa je korpus slWaC). Ciljni pristop je primernejši za izdelavo manjših zbirk besedil, pri katerih so zaradi specifik raziskav še posebej pomembni struktura in metapodatki besedil (primer tega pristopa je v prejšnjem razdelku predstavljeni korpus JANES).

Pajkanje se tipično začne z začetnim naborom spletnih dokumentov, nato pa pajek s pomočjo povezav v dokumentih zbira nove dokumente. Vprašanje, ki se tu postavlja, je, kako omejiti nabor zajetih dokumentov, da ne bodo bodisi v napačnem jeziku ali napačne zvrsti glede na namene pajkanja. Razlikujemo dva osnovna pristopa k določanju dokumentov za pajkanje. Prvi temelji na *omejitvah naslovov URL*, npr. na vrhnji domeni ».si« ali na »med.over.net«, drugi pa na *seznamu ključnih besed*, ki definirajo ciljno področje diskurza, kot je npr. okolje, turizem, kulinarika itd. Za slednje je tipično, da se za zajem dejanskih naslovov URL za pajkanje uporablja katerega od programskih vmesnikov (API) spletnih iskalnikov. Večina pajkanja za splošne korpuse spletnih besedil se izvaja s pomočjo omejitev naslovov URL (na ta način je pajkanje potekalo tudi pri obstoječi Gigafidi), za specializirane korpuse pa je primernejša izbira s pomočjo seznama ključnih besed, pri čemer sta bolj znani orodji za ta pristop BootCaT (Baroni in Bernardini 2004) in WebBootCaT (Baroni et al. 2005).

Na spletu najdemo dokumente v več različnih formatih. Najpogostejši so dokumenti HTML, ki so za zajem v korpus problematični zato, ker je velik del njihove vsebine lahko namenjen izgledu, dostikrat pa vsebujejo tudi ponavljajoče se dele, ki za besedilni korpus predstavljajo šum. Drugi format dokumentov, ki tudi vsebujejo veliko jezikoslovno zanimivih podatkov, vendar se jih bistveno redkeje zajema in obdeluje, pa so dokumenti v formatu PDF. Problem pri zajemu besedil iz takih dokumentov je, da je format PDF namenjen tiskanju, zato ni kodiran kot niz znakov, temveč kot nabor znakov s svojim položajem na posamezni strani. V nadaljevanju se zato osredotočamo predvsem na opis obdelave dokumentov HTML (za dokumente PDF bi bilo namreč treba prilagoditi proces luščenja vsebine dokumentov).

Posebej je treba omeniti še spletne platforme, na katerih, vsaj izvorno, besedila niso postavljena na splet kot dokumenti HTML (ali PDF), pač pa so prejemnikom poslana kot posamezna sporočila, podobno kot SMS-sporočila ali elektronska pošta. Tu je daleč najbolj znana platforma Twitter, sistem, ki omogoča pošiljanje krajših sporočil sledilcem. Twitter ponuja tudi programske vtičnike API, ki omogočajo zajemanje tvitov posameznikov ali tem. Kot je bilo prikazano zgoraj, smo v korpus JANES vključili tvite, ki smo jih zbrali z orodjem TweetCat (Ljubešić et al. 2014), ki je bilo namensko izdelano za gradnjo korpusov tvitov manjših jezikov. To orodje s pomočjo začetnega seznama specifično slovenskih besed identificira uporabnike, ki tvitajo pretežno v slovenščini, nato pa prek prijateljev in sledilcev postopoma širi nabor uporabnikov ter zbira njihove tvite skupaj z metapodatki.

5.1 Postopki pri generičnem pajkanju

Kot rečeno, se generično pajkanje uporablja predvsem takrat, kadar je cilj pridobiti velike količine besedil (z več kot milijardo pojavnic) ali pa so človeški viri za pridobivanje podatkov omejeni.

Osnovnih korakov, ki se izvajajo pri generičnem pajkanju jezikoslovno relevantnih podatkov in jih uporablja tudi sistem, s katerim smo zgradili korpus slWaC, je več. Prvi korak je *izdelava seznama spletnih strani*, ki so izhodišče za pajkanje. V primeru jezikov z manjšim številom govorcev, kakršna je slovenščina, je to večinoma nekaj bolj znanih spletnih strani v izbranem jeziku. Drugi korak je *pajkanje*, ki se, tehnično gledano, tipično izvaja v večnitnem načinu in s pregledovanjem povezav v širino, pri čemer se seznam strani, ki jih je treba zajeti, sproti dopolnjuje z identifikacijo povezav z že zbranih spletnih strani. Ko je dokument zajet, je treba najprej ugotoviti, kateri *kodni nabor znakov* uporablja. Ta podatek naj bi bil sicer zapisan v metapodatkih dokumenta HTML, vendar v praksi dostikrat manjka ali ni pravilen glede na dejansko kodiranje dokumenta. Ugotavljanje kodnega nabora zato večinoma poteka na podlagi primerjave distribucije bajtov v besedilu dokumenta z distribucijo vnaprej pripravljenih dokumentov z znanimi kodiranji.

Pri generičnem pajkanju ne obstaja vnaprej predvidljiva predloga videza dokumenta, zato je treba za *zajem vsebine* uporabiti splošne programe, kot so jusText (Pomikálek 2011) in Boilerpipe (Kohlschütter et al. 2010). Ta korak, ki zaradi svoje generičnosti dokument strukturira največ do obsega odstavka, ne zajame metapodatkov o besedilu in tipično tudi ne odstrani vsega nebesedilnega šuma iz dokumenta. Na osnovi zajete besedilne vsebine dokumenta je nato treba *identificirati jezik* dokumenta. Splet je namreč večjezično okolje, zato je ta korak nujen pri izdelavi korpusa. Orodje, ki daje pri tem dobre rezultate, je program langid.py (Lui in Baldwin 2012), napisan v programskem jeziku Python. Zadnji

korak je *odstranjevanje (približnih) dvojnikov*, saj se enaka ali skoraj enaka besedilna vsebina pogosto pojavlja na več različnih naslovih URL. Postopki odstranjevanja (približnih) dvojnikov večinoma temeljijo na računanju preseka n-gramov besed med dvema dokumentoma. Tipična hevristika je, da če se 7-grami dveh dokumentov prekrivajo v več kot polovici primerov, lahko enega od njiju odstranimo kot približnega dvojnika.

Opisanih šest korakov se večinoma izvaja ločeno, zaradi česar je postopek pajkanja daleč od optimalnega. Izjema je SpiderLing (Suchomel in Pomikálek 2012), ki združuje korake od pajkanja do identifikacije jezika v integriran postopek, pri katerem posamezni koraki medsebojno komunicirajo s ciljem optimizacije količine prevzetih podatkov in velikosti končnega korpusa.

5.2 Postopki pri ciljnem pajkanju

Ciljno pajkanje se uporablja, kadar se pajka manjša količina podatkov oz. kadar je človeških virov za izvedbo postopkov dovolj. Takšno pajkanje ima tri osnovne korake. Specializirani korpusi se najpogosteje gradijo na osnovi določene vsebine, ne na osnovi spletnih domen. Prvi korak je zato *identificiranje spletnih domen* oz. njihovih delov, za katere se predvideva, da so bogati z želenimi vsebinami. Pri tem je treba upoštevati tudi tehnično-pravne omejitve posameznih izvorov, npr. ali spletno mesto prepoveduje pajkanje (datoteka robots.txt), ali izvor ponuja programski vmesnik API za zajem podatkov (npr. Twitter) in ali morda celo omogoča prevzem celotne baze besedil (npr. Wikipedija). To dvoje izrazito olajša zajem, medtem ko uporaba tehnologij, kot so pošiljanje POST, AJAX ipd., zelo oteži izdelavo ekstraktorjev. Naslednji korak je *pajkanje*, ki večinoma zajame vse oz. čim več dokumentov z izbranih domen. Najbolj zahtevno in zamudno je pisanje *ekstraktorjev*, v katerih mora programer opisati shemo HTML-dokumentov za vsak posamezen vir, pri čemer je lahko struktura dokumentov zelo kompleksna, npr. pri zajemu časopisnih prispevkov, pri katerih bi hoteli zajeti tudi zaporedje komentarjev vsakega prispevka.

5.3 Spremljevalni korpusi

Splet je izrazito naklonjen gradnji spremljevalnih korpusov, saj se njegova vsebina nenehno spreminja in dopolnjuje, pri čemer je potem, ko je sistem pajkanja postavljen, ponovno zbiranje podatkov ter beleženje razlik oz. povsem novih dokumentov enostavno. To velja tako za generično kot za ciljno pajkanje, pri čemer je generični pristop robustnejši, saj lahko posamezni izvori spremenijo obliko svojih strani, s čimer stari ciljni ekstraktorji prenehajo pravilno delovati.

Najboljši pri sprotnem zajemanju spleta so spletni iskalniki, predvsem Google, pa tudi lokalni, kot je Najdi.si, saj ti nepretrgoma in intenzivno pregledujejo splet ter na njem iščejo nova besedila. Težko si sicer zamislimo, da bi za jezikoslovne potrebe lahko uporabljali tako intenzivno pajkanje, nam pa lahko služi vsaj kot zgornja meja možnega. Od posameznega projekta je odvisno, kako se bodo tako glede na svoje potrebe kot tudi zmožnosti sodelujoči raziskovalci odločili glede pogostosti ponovnega pajkanja izbranih vsebin. Za leksikografske namene bi bil spremljevalni korpus, ki bi nastajal sproti vse od prve faze projekta dalje, gotovo dragocen za zaznavanje večjih in nenadnih leksikalnih sprememb, ki jih povzročajo dogodki ter pojavi, o katerih v nekem obdobju bolj intenzivno poročajo mediji in posledično vzbujajo tudi večje zanimanje govorcev – potencialnih uporabnikov slovarja. Potem ko je prva različica slovarja narejena in že na voljo, slovar pa bi hoteli sproti vsebinsko posodabljati, pa postane izdelava spremljevalnega korpusa in metod za ugotavljanje nove leksike, pomenskih sprememb ali sprememb značilnega besedilnega okolja še veliko bolj pomembna, pravzaprav ključna.

6 SKLEP

V sodobnem jezikoslovju so paradigme, ki na rabo nestandardnih jezikovnih različic v internetnem pisnem komuniciranju gledajo kot na odraz nepopolnosti ali osiromašenosti sporazumevalnih zmožnosti, preživete, saj analize jezikovne rabe na internetu ugotavljajo sposobnost uporabnikov, da se prilagodijo računalniškemu mediju oz. da zmožnosti medija izrabijo za zadovoljevanje svojih komunikacijskih potreb, da si prizadevajo skrajšati in poenostaviti pisanje, predvsem pa da pisanje približajo svoji identiteti ter govoru (Herring 2001).

Internetno komuniciranje danes tvorno dopolnjuje in spreminja podobo javne pisne slovenščine do mere, ki je sodobni slovar gradivno ne more več zaobiti. V prispevku smo skušali prikazati, kako je mogoče spletni del Gigafide nadgraditi tako v obsegovnem kot v tematskem in žanrskem smislu, ter opozorili, da ga je treba v korpus umestiti na razviden način (tj. z bolj razdelanimi taksonomskimi kategorijami). Del spletnih žanrov je zapisan v nestandardni slovenščini, kar v korpusno jezikoslovje prinaša še dodaten jezikovnotehnološki izziv: premostitev ovir za njihovo avtomatsko obdelavo, vendar pa viri in orodja, s katerimi si je mogoče pri tem pomagati, v slovenskem prostoru že nastajajo, preizkušene pa so bile tudi že različne metode pajkanja. Namen na predlagani način nadgrajene Gigafide je torej zajeti javnosti namenjeno pisno produkcijo spletne slovenščine v širokem smislu; izbor in interpretacija podatkov iz takega korpusa za potrebe slovarja pa bosta nato prepuščena odločitvam akterjev naslednje faze tega procesa – leksikografom.

Jezikovne tehnologije in zapis korpusa

Tomaž Erjavec, Peter Holozan in Nikola Ljubešić

Abstract
This paper provides an overview of the levels of primarily automatic linguistic annotation that should be part of the annotation of corpora to be used inter alia as the basis for lexicographic analyses of contemporary Slovenian language. An overview of existing research in this field is outlined, with the focus then turning to a specific set of open source and mainly language independent tools and their models for Slovenian, with suggestions provided for their improvement. A short description of the proposed corpus encoding is also presented.

Keywords: linguistic annotation, corpora, annotation format

Ključne besede: jezikoslovno označevanje, korpusi, zapis oznak

1 UVOD

Prispevek obravnava jezikoslovne oznake, zapisane v korpusih, ki bi lahko služile kot osnova slovaropisnim dejavnostim, in format zapisa teh korpusnih oznak. Jezikovnotehnološke oznake se v korpuse zapišejo avtomatsko s pomočjo orodij, napisanih ali naučenih za označevanje slovenskega jezika, vendar pa ni izključeno, da se naknadno ne popravljajo, bodisi ročno ali pa avtomatsko, s ciljnimi ekspertnimi programi.

Prispevek ne pokriva vseh jezikovnotehnoloških orodij, ki jih je potrebno ali koristno imeti pri slovaropisnem delu s korpusi, temveč samo tista, za katera se predvideva, da bodo njihovi izhodi, torej oznake, dejansko našli mesto v zapisu korpusa. Zato tu ne obravnavamo orodij, namenjenih luščenju informacij iz korpusov, pač pa se osredotočamo na ravni označevanja, ki lahko služijo kot vir znanja o jeziku za luščilne programe, od konkordančnikov do programov za luščenje sopomenk. Ob tem se tudi večinoma omejimo na programe, ki so že bili razviti za slovenski jezik. Obravnavali bomo naslednje ravni označevanja, ki se večinoma tudi izvajajo v podanem zaporedju:

1. **tokenizacija**, ki razdeli besedilo na posamezne pojavnice, bodisi besede ali ločila; ta korak tipično identificira posebne razrede pojavnic, kot so cifre, okrajšave, naslove URL itd., vanj pa tipično sodi tudi **segmentacija** besedila na povedi, saj je ta neposredno odvisna od tokenizacije;
2. **normalizacija**, ki se uporablja za pretvorbo nestandardnih besednih oblik (kot jih najdemo npr. v uporabniško ustvarjenih besedilih na spletu) v standardne z namenom lažjega iskanja besed, kot tudi zato, da lahko nad normaliziranimi besedami uporabljamo standardna orodja za naslednje korake označevanja;
3. **oblikoskladenjsko označevanje**, ki besednim pojavnicam v besedilu pripiše sobesedilno odvisno enolično oblikoskladenjsko oznako, kot npr. »obči samostalnik moškega spola v imenovalniku dvojine«;
4. **lematizacija**, ki besedni obliki, tipično glede na njeno oblikoskladenjsko oznako, pripiše njeno osnovno obliko;
5. **skladenjsko razčlenjevanje**, ki vsako poved v besedilu skladenjsko razčleni; ta raven označevanja za slovaropisje sicer ni nujna, lahko pa se izkaže za koristno, še posebej pri proučevanju besednih zvez.

Obstajajo še druge ravni označevanja, ki so tudi lahko koristne, a niso nujne za slovaropisje, hkrati pa jih je težje umestiti v zaporedje zgoraj podanih korakov, saj jih lahko, odvisno od pristopa, določamo na podlagi neobdelanega besedila,

pri čemer jih umestimo neposredno po koraku 1 oz. 2, ali pa uporabimo več informacij, kar pomeni, da jih izvajamo po koraku 4 ali celo 5. Od teh orodij so bila nekatera sicer že razvita za slovenski jezik, čeprav samo prototipno, razvoj drugih pa je še povsem v zametkih. Ta označevanja, katerih rezultati se tudi lahko zapišejo v korpus, so:

6. **določanje imenskih entitet**, ki v besedilu identificira lastna imena in jih klasificira, npr. v osebna imena, zemljepisna imena, imena podjetij in ustanov itd.; ob tem nekateri sistemi identificirajo še številske in druge izraze ter jih klasificirajo, npr. v denarne enote, datume itd.;
7. **določanje terminov**, pri čemer se je treba zavedati, da je definicija termina razmeroma problematična, saj koncept termina ni vedno enoumno določen, temveč je pogosto odvisen od področja, ciljne publike ipd.;
8. **pripisovanje semantičnih informacij**, kjer besedam ali besednim zvezam pripišemo njihov pomen glede na neki semantičnoleksikalni vir, lahko pa jih tudi medsebojno povežemo s semantičnimi vlogami; čeprav bi takšno označevanje bilo izredno koristno za leksikografijo, zaradi kompleksnosti tega problema trenutni programi verjetno niso dovolj natančni, da bi dajali res uporabne rezultate.

Za orodja, ki opravljajo našteta označevanja, kot tudi za jezikovnotehnološka orodja na splošno tipično uporabljamo dva načina prilagajanja na določen jezik:

- Orodja uporabijo ročno napisana pravila, ki sicer zahtevajo veliko človeškega dela, lahko pa dajo (odvisno od ravni označevanja) zelo dobre rezultate. Takšna orodja se dostikrat uporabljajo za segmentacijo besedila, npr. v pojavnice ali termine, ali pa, tradicionalno, za morfološko analizo. Pri označevanjih, kot sta oblikoskladenjsko označevanje in skladnja, je teh pravil zelo veliko, pri čemer so dostikrat tudi medsebojno odvisna, kar zelo otežuje njihov razvoj in razhroščevanje v primerih, ko ne dajo želenih rezultatov.
- Orodja se naučijo modela določenega jezika na osnovi učnih podatkov, torej (ročno) označenih korpusov ali drugih jezikovnih virov. Metode strojnega učenja se hitro razvijajo, vendar pa potrebujemo za izdelavo kvalitetnih modelov čim večje označene jezikovne vire – njihova izdelava pa je tipično zamudna in draga.

Obe vrsti orodij tipično uporabljata tudi zaledne vire znanja o jeziku, predvsem leksikon (Dobrovoljc et al. 2015).

2 PREGLED ORODIJ ZA SLOVENSKI JEZIK

Za vse naštete ravni označevanja so bila za slovenski jezik že razvita (vsaj prototipna) orodja. V nadaljevanju razdelka podajamo pregled glavnih, ki jim je skupno to, da se vsaj do neke mere še vedno vzdržujejo in so tipično tudi prosto oz. odprto dostopna. Zato med našteta orodja ne vključimo sicer enega prvih sistemov za oblikoskladenjsko označevanje slovenskega jezika, ki je bil razvit na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU za potrebe označevanja njihovih korpusov (Jakopin in Bizjak Končar 1997).

2.1 Orodja podjetja Amebis

Orodja podjetja Amebis po eni strani niso odprto dostopna, po drugi pa predstavljajo pri nas najdlje razvijan integriran sistem označevalnih orodij in zalednih virov, ne samo prilagojenih, temveč napisanih posebej za obdelavo slovenskega jezika. Orodja so bila v osnovi razvita za slovnični pregledovalnik Besana (Holozan 2012) in strojni prevajalnik Presis (Romih in Holozan 2002), napisana so v programskem jeziku C++, delujejo v 32- in 64-bitni različici. Njihova struktura nekoliko odstopa od klasične: tudi tukaj je na prvem mestu tokenizacija, njena pomembna lastnost pa je, da kot enoto obravnava tudi posebne pojavnice, kot so spletni in elektronski naslovi, telefonske številke ter smeški. V drugem koraku sledi označevalnik, ki besedam s pomočjo slovarja pripiše vse možne kombinacije lem in oblikoskladenjskih oznak (slovar trenutno vsebuje več kot 7,6 milijona elementov). Tudi označevalnik prepoznava pojavnice, kot so spletni naslovi, kemijske spojine in smeški, hkrati pa išče še morebitne zatipkane besede in nekatere tipične nestandardne oblike, pri čemer je del nestandardnih besed že v slovarju s posebnimi oblikoskladenjskimi oznakami. Zadnji korak je analizator, ki za vsako besedo izbere najverjetnejši par leme in oblikoskladenjske oznake, hkrati pa v Amebisovem vmesnem jeziku (Holozan 2011) zapiše tudi skladenjsko razčlembo in pomene besed, ki so vzeti iz baze Ases (Arhar in Holozan 2009). Analizator lahko po potrebi vpliva tudi na tokenizacijo oz. segmentacijo besedila, npr. pri razreševanju primerov tipa »Videl sem ga. Micka ga je tudi videla.« (v primerjavi s »Prišla je še ga. Micka.«). Tu bi naivni tokenizator »ga.« lahko identificiral kot eno pojavnico (okrajšavo) in besedilo kot eno poved, Besana pa pravilno kot dve pojavnici in dve povedi. Amebisova orodja delujejo s pomočjo ročno napisanih pravil in podatkov v bazi Ases, pri čemer je najpomembnejši koncept glagolskih predlog (Holozan 2011), ki vsebujejo podatke o vezljivosti glagolov, vnesenih pa je tudi veliko lastnih imen, razporejenih v približno 30 kategorij (kar omogoča tudi določanje imenskih entitet). Narejen je bil tudi program, ki besedilo tokenizira,

lematizira in oblikoskladenjsko označi v skladu s specifikacijami, ki so bile uporabljene v projektu Sporazumevanje v slovenskem jeziku in jih implementira tudi označevalnik Obeliks (o tem več v nadaljevanju).

Z Amebisovimi orodji sta bila med drugim označena korpusa Fida in FidaPLUS. Za širšo uporabo Amebisovih orodij je največja prepreka to, da so lastniška in kot taka niso odprtokodna: za njihovo uporabo je, vsaj trenutno, potrebno skleniti dogovor z Amebis, d. o. o.

2.2 Označevalnik To(Tr)TaLe

Na Odseku za tehnologije znanja IJS je bilo v zaporedju več projektov razvito orodje ToTaLe (Erjavec et al. 2005), ki implementira cevovod treh modulov: tokenizatorja, ki besedilo tudi segmentira na povedi, oblikoskladenjskega označevalnika in lematizatorja. Za tokenizacijo ToTaLe uporablja modul mlToken, večjezični tokenizator, ki za prilagoditev na posamezni jezik uporablja jezikovno odvisne sezname, npr. okrajšav ali pravil za zapis števil. ToTaLe za oblikoskladenjsko označevanje uporablja program TnT (Brants 2000), zdaj že razmeroma star označevalnik, ki ga naučimo modela posameznega jezika nad ročno označenim korpusom, lahko pa uporablja tudi zaledni leksikon. Trenutni model je bil naučen na korpusu jos1M (Erjavec et al. 2010; Erjavec in Krek 2010), za zaledni leksikon pa uporablja kar besedišče korpusa FidaPLUS (Arhar in Gorjanc 2007). Za lematizacijo ToTaLe uporablja program CLOG (Erjavec in Džeroski 2004), ki na osnovi besedne oblike in njene oblikoskladenjske oznake besedni obliki določi njeno osnovno obliko. Tudi ta program se nauči modela lematizacije avtomatsko, na osnovi učne množice, ki je v tem primeru seznam trojčkov (besedna oblika, oblikoskladenjska oznaka, lema). Seznam, ki je programu služil kot učna množica, je bil zajet iz kombinacije besedišča korpusa jos100k, ročno pregledanih pojavnic v korpusu jos1M in izbranih besed iz korpusa FidaPLUS. ToTaLe je dostopen za uporabo na spletu, z njim pa je bila označena tudi večina korpusov, ki so dostopni za pregledovanje na konkordančniku noSketchEngine (Rychly 2007), instaliranem na nl.ijs.si (Erjavec 2013).

ToTaLe je napisan v programskem jeziku Perl, ravno tako tudi modula za tokenizacijo in lematizacijo. Perl sicer dandanes ni več zelo popularen programski jezik, vseeno pa obstajajo zanj implementacije za vse glavne operacijske sisteme. Zato pa oblikoskladenjski označevalnik TnT, ki ga uporablja ToTaLe, ni odprtokoden in je dostopen samo za nekomercialno rabo, in to samo v prevedeni obliki za operacijski sistem Linux, tako da ToTaLe v trenutni postavitvi ne more biti niti odprto dostopen niti ni uporaben npr. na operacijskem sistemu Windows.

V nadaljevanju raziskav je bilo razvito orodje ToTrTaLe, ki je enako kot ToTaLe, a z dvema novostma: po modulu za tokenizacijo smo vključili tudi (opcijske) module za transkripcijo, za razliko od orodja ToTaLe – ki pričakuje kot vhod golo besedilo in izhod vrne v obliki tabelarične datoteke – pa ToTrTaLe na vhodu pričakuje datoteko XML, ki uporablja shemo TEI, ravno tako pa vrne izhod v obliki TEI XML. Transkripcijski modul je mišljen za uporabo pri starejših (slovenskih) besedilih z namenom posodobitve posameznih besed, s čimer bistveno olajša delo oblikoskladenjskemu označevalniku in lematizatorju, saj sta oba naučena za obdelavo sodobnih besedil. Za posodabljanje modul za transkripcijo trenutno uporablja orodje Vaam (Reffle 2011) z ročno napisanimi pravili za posodabljanje starejših slovenskih besed. Z orodjem ToTrTaLe je bil zaenkrat označen samo korpus starejše slovenščine IMP (Erjavec 2015).

Obe orodji sicer dajeta razmeroma dobre rezultate, skupno pa jima je to, da nista najbolje vzdrževani: bilo bi ju npr. dobro na novo naučiti modelov za slovenski jezik, saj so sedaj na voljo boljši viri, predvsem leksikon Sloleks (Arhar 2009; Dobrovoljc et al. 2013) in korpus ssj500k. Tudi sicer je okolje, v katerem sta narejeni, sedaj že zelo zastarelo, kar velja tudi za posamezne module programov. Vsaj TnT bi bilo nujno potrebno zamenjati s kakšnim sodobnejšim, predvsem pa odprtokodnim in od operacijskega sistema neodvisnim označevalnikom.

Kot omenjeno, je bila normalizacija v kontekstu posodabljanja starejših besed slovenskega jezika v sklopu orodja ToTrTaLe že implementirana. Vendar so bila pravila tam napisana ročno, izkaže pa se, da boljše rezultate dajejo avtomatsko naučeni modeli normalizacije, ki za osnovo uporabljajo statistično strojno prevajanje na osnovi znakov (Scherrer in Erjavec 2013). Osnovno orodje, ki se uporablja za takšno normalizacijo, je statistični strojni prevajalnik Moses (Koehn et al. 2007), ki ga izšolamo na parih izvorna (nestandardna) beseda : normalizirana beseda. Ta pristop je uporaben ne samo za posodabljanje starejših besed, temveč tudi za predobdelavo sodobnih, nestandardno zapisanih besedil, kakršne npr. najdemo v spletnem komuniciranju, zaradi česar je relevanten tudi za slovar sodobnega slovenskega jezika. Pristop s statističnim strojnim prevajanjem na osnovi znakov je že bil preizkušen za standardizacijo besed v slovenskih tvitih (Ljubešić et al. 2014), in to s spodbudnimi rezultati. Pri normalizaciji besed se seveda pojavi tudi vprašanje, katera besedila normalizirati; če namreč orodje uporabljamo tudi pri standardnih besedilih, je velika verjetnost, da bo »normaliziralo« tudi povsem standardne besede, s čimer bo naredilo več škode kot koristi. Tudi tu so bile že izvedene začetne raziskave, pri katerih se je sistem regresijskega strojnega učenja na osnovi manjše, ročno označene množice slovenskih tvitov in drugih uporabniško generiranih vsebin s spleta naučil pripisati stopnjo nestandardnosti novim besedilom (Ljubešić et al. 2015). Normalizacijo bi tako lahko uporabili samo pri besedilih, ki so avtomatsko označena kot nestandardna.

2.3 Označevalnik Obeliks in skladenjski razčlenjevalnik za slovenščino

V okviru projekta Sporazumevanje v slovenskem jeziku je bilo razvito orodje Obeliks (Grčar et al. 2012), ki vhodno besedilo tokenizira, segmentira na povedi, oblikoskladenjsko označi in lematizira. Za tokenizacijo Obeliks uporablja modul z ročno napisanimi pravili, za oblikoskladenjsko označevanje namensko razvit modul strojnega učenja po sistemu največje entropije, za lematizacijo pa sistem LemmaGen (Juršič et al. 2010), ki prav tako temelji na strojnem učenju. Posebnost oblikoskladenjskega označevalnika je ta, da za označevanje ne uporablja samo modela, avtomatsko naučenega iz učnega korpusa, pač pa tudi ročno napisana ekspertna pravila, ki filtrirajo hipoteze, ki jih je generiral model, poleg tega pa usklajuje rezultate lematizatorja in označevalnika, tako da ne prideta v kontradikcijo. Obeliks je bil naučen na ročno označenem korpusu ssj500k (Arhar 2009; Krek et al. 2013) in izmed javno dostopnih orodij dosega najboljše rezultate za slovenski jezik. Trenutno največji problem tega označevalnika je verjetno ta, da je implementiran v programskem jeziku C#, ki je namenjen sistemom Windows, kar pomeni, da program ni enostavno prenosljiv na druge platforme, kot je Linux.

Z Obeliksom so bili označeni korpus govorjene slovenščine Gos (Verdonik in Zwitter Vitez 2011), korpus besedil odnosov z javnostmi KoRP (Logar 2013), korpus šolskih pisnih izdelkov Šolar (Rozman et al. 2012) in korpus Gigafida; oznake iz Gigafide pa so (skupaj z besedili) prisotne še v korpusih KRES, ccGigafida in ccKRES (Erjavec in Logar 2012).

V okviru projekta Sporazumevanje v slovenskem jeziku je bil razvit tudi skladenjski razčlenjevalnik za slovenščino (Dobrovoljc et al. 2012) oz. natančneje: odvisnostnoskladenjski razčlenjevalnik MSTParser (McDonald et al. 2006) je bil naučen označevanja skladenjskih drevesnic na korpusu ssj500k. Razčlenjevalnik daje razmeroma dobre rezultate, je pa, kot je to tudi sicer običajno pri kakršnemkoli jezikoslovnem označevanju oz. razčlenjevanju, točnost zelo odvisna od zvrsti besedila – bolj ko zvrst besedila za označevanje odstopa od zvrsti učne množice, slabši so rezultati. Pri evalvaciji razčlenjevalnika se je izkazalo tudi to, da je točnost zelo odvisna od vrste odvisnostne povezave, saj sega od 54 do 96 %.

2.4 Druga orodja

Za označevanje imenskih entitet (NER) sta bili za slovenščino razviti dve orodji. Na IJS je bil razvit označevalec NER, ki uporablja strojno učenje na osnovi

pogojnih naključnih polj (Štajner et al. 2012), in to z modelom, naučenim na korpusu ssj500k. Program je sicer objavljen pod odprtokodno licenco, sta pa njegova namestitev in način uporabe razmeroma slabo dokumentirana, kar otežuje njegovo uporabo. V drugi raziskavi (Ljubešić et al. 2013) je bilo uporabljeno orodje StanfordNER (Finkel et al. 2005), ki ravno tako deluje na osnovi pogojnih naključnih polj. Tudi tu je bil model naučen na korpusu ssj500k, vendar v kombinaciji s korpusom spletne slovenščine slWaC (Ljubešić in Erjavec 2011). Ta omogoča boljši zajem distribucijskih značilk, ki se izkažejo za zelo koristne pri zmanjšanju števila napak, kot tudi izboljšanju priklica. Modeli za program so odprto dostopni, StanfordNER pa je vzdrževano in dokumentirano orodje.

Za luščenje terminov so bili za slovenski jezik opravljeni številni eksperimenti in izdelana mnoga orodja (Logar in Vintar 2008; Vintar 2009; 2010; Logar et al. 2013), vendar slednja niso odprto dostopna ali vzdrževana. Orodja temeljijo predvsem na kombinaciji jezikoslovnega védenja o terminih (predvsem nizih oblikoskladenjskih oznak, ki lahko predstavljajo termine) ter izrabi matematičnih lastnosti porazdelitve besed in besednih nizov v korpusih. Vsaj za slovenščino pri identifikaciji terminov še niso bile uporabljene metode strojnega učenja, obenem pa tudi ne obstajajo odprte učne množice, ki bi jih lahko za to uporabili.

3 SMERNICE ZA NADALJNJI RAZVOJ OZNAČEVANJA KORPUSOV

3.1 Izboljšanje označevalnih shem

Preden se posvetimo izboljšanju samih orodij oz. korpusov, na katerih se orodja učijo, je smiselno odpreti vprašanje označevalnih shem, na osnovi katerih so korpusi (ročno) označeni. Zasnovo teh shem bi bilo namreč koristno na novo premisliti in izvesti testiranja, kar bi povečalo točnost orodij, obenem pa ohranilo ali celo izboljšalo jezikoslovno informativnost kategorij posameznih ravni označevanja.

Slovnične informacije o posameznih besedah v korpusih ssj500k, Gigafida, KRES itd., kot tudi v oblikoskladenjskem leksikonu Sloleks, temeljijo na oblikoskladenjskih specifikacijah, razvitih v okviru projekta Jezikoslovno označevanje slovenščine JOS (Erjavec in Krek 2008). Ta sistem ima svoj izvor in je usklajen s specifikacijami MULTEXT (Ide in Véronis 1994) oz. MULTEXT-East, pri čemer so specifikacije MULTEXT-East 4.0 (Erjavec 2012) za slovenščino identične specifikacijam JOS in pokrivajo 12 jezikov, od tega skoraj vse slovanske.

Specifikacije JOS definirajo 12 besednih vrst: samostalnik, pridevnik, glagol, prislov, zaimek, števnik, predlog, veznik, členek, medmet, okrajšavo in neuvrščeno.

Večina besednih vrst ima pripisane oblikoskladenjske lastnosti, kot so npr. pri samostalniku vrsta (obči, lastni) ali sklon. Možne kombinacije besedne vrste in njihovih lastnosti so kodirane kot nizi, v katerih vsaki poziciji v nizu ustreza en atribut, enočrkovna koda pa poda njegovo vrednost. Tako npr. niz Somei pomeni besedna vrsta = samostalnik, vrsta = občno ime, spol = moški, število = ednina, sklon = imenovalnik. Kode v nizu in tudi lastnosti (torej atributi in njihove vrednosti) so prevedene v angleški jezik, da olajšajo uporabo sistema v mednarodnem okolju. Tako je npr. Somei po angleško Ncmsm oziroma Category = Noun, Type = common, gender = masculine, number = singular, case = nominative. Sistem JOS skupaj loči 1.902 različni oznaki, ki so v specifikacijah naštete, razstavljene po lastnostih in ilustrirane s primeri iz korpusa. Oznake JOS, kot je Somei, se uporabljajo v oblikoskladenjsko označenih korpusih, kjer služijo kot učna množica za učenje oblikoskladenjskih označevalnikov, s katerimi nato avtomatsko označujemo nove korpuse. Te oznake se uporabljajo tudi v oblikoskladenjskem leksikonu Sloleks, saj z njimi definiramo besedne oblike v paradigmi posameznih besed.

Oznake JOS se sicer res uporabljajo v velikem številu korpusov, vendar ni nujno, da je ta nabor oznak oz. lastnosti najprimernejši za vse aplikacije. Tako si lahko zamislimo nabor oznak, ki izpusti vse lastnosti, pri katerih se oblikoskladenjski označevalniki najbolj motijo (npr. sklon), ali pa vse pregibne lastnosti, če je za namene projekta dovolj besedam pripisati samo njihove leksikalne lastnosti. Takšne alternative, ki zmanjšajo velik nabor oznak JOS in s tem povečajo točnost označevalnikov, so se že pojavile: poglobljena študija v Krek (2011) predlaga več možnih redukcij oznak, medtem ko Erjavec (2015) omejuje nabor na 32 oznak, ki zajemajo samo besedno vrsto in nekaj njihovih leksikalnih lastnosti. Vseeno bi bilo potrebnih več raziskav, ki bi opredelile optimalen nabor oznak za posamezne potrebe.

V zadnjem času se je pojavila še ena zanimiva možnost, in sicer projekt Universal Dependencies (Nivre et al. 2015), v katerem se izdelujejo specifikacije in drevesnice za veliko število jezikov, mdr. tudi za slovenščino. Ob definiciji skladenjskih povezav prinaša projekt tudi univerzalen nabor oblikoskladenjskih lastnosti (z možnostjo jezikovnospecifičnih razširitev). Kljub temu, da težnja k univerzalnosti nujno privede do slabše prilagojenosti sheme za posamezni jezik, pa dejstvo, da je shema nato primerljiva z mnogimi drugimi jeziki, morda vseeno odtehta posamezne slabše rešitve.

Še bolj kot oblikoskladenjska raven označevanja potrebuje nadaljnje raziskave manj preizkušen nabor skladenjskih oznak in povezav projektov JOS ter Sporazumevanje v slovenskem jeziku (Erjavec et al. 2010; Arhar 2009), pri katerem npr. množično povezovanje pojavnic na koren drevesa predstavlja problem pri izdelavi

avtomatskih skladenjskih razčlenjevalnikov (Javoršek 2015). Tudi pri nadaljnjem razvoju skladenjskega sistema označevanj bi bilo zelo koristno upoštevati priporočila in prakso projekta Universal Dependencies.

Pomanjkljivost trenutnih kategorij se je pokazala tudi pri označevanju imenskih entitet, saj so raziskave (Štajner et al. 2012) pokazale, da z razdelitvijo »ostalih« imenskih entitet (torej tistih, ki niso osebna ali zemljepisna imena) na imena organizacij in »ostale« v učnem korpusu ssj500k dosežemo boljše rezultate prepoznavanja tudi pri drugih razredih. Z uvedbo nove kategorije smo v tem primeru ne samo obogatili nabor označevalnih kategorij, temveč tudi pripomogli h kvaliteti označevanja. To se sklada z ugotovitvami avtorjev češkega korpusa imenskih entitet CNCEC, pri katerem ločijo kar 62 kategorij imenskih entitet (Ševčíková et al. 2007). Avtorji so ugotovili tudi, da zmanjšanje števila kategorij imenskih entitet privede do slabših rezultatov označevanja. Tudi za slovenski jezik bi zato kazalo razmisliti o nadaljnjem povečanju števila kategorij za imenske entitete.

3.2 Izboljšanje točnosti orodij

Opisana orodja se med seboj zelo razlikujejo tako po kvaliteti označevanja kot po enostavnosti uporabe. Pri vseh bi bilo seveda koristno izboljšati točnost označevanja, saj je vsaka napaka problematična z dveh vidikov. Po eni strani predstavlja šum v podatkih, saj slovaropisec s svojo poizvedbo dobi tudi rezultate, ki so napačni. Tako npr. pri poizvedovanju po določeni lemi oz. osnovni obliki lahko besede, ki jim je bila pri lematizaciji pripisana napačna skupna osnovna oblika, zameglijo sliko njenih konkordanc, kolokacij, besednih skic itd. Vendar tu slovaropisec lahko vsaj pregleda primere in se odloči, kateri so pravilni, kar je sicer zamudno, a izvedljivo. Toliko bolj nevaren pa je pomanjkljiv priklic orodij, saj v tem primeru nekaterih podatkov slovaropisec enostavno ne vidi, ker mu jih orodja ne odkrijejo: če je neka beseda popolnoma ali večinoma narobe lematizirana, je z iskanjem po njeni lemi ne bomo našli ali bo ponujenih zelo malo zadetkov. Glavni cilj pri avtomatskem označevanju korpusov bi zato morala biti težnja po izboljšanju točnosti in priklica vseh obravnavanih metod – še zlasti v njihovih prvih korakih (tokenizacija, oblikoskladenjsko označevanje, lematizacija), saj vsaka napaka množi napake vseh nadaljnjih stopenj obdelave in s tem tudi otežuje leksikografsko analizo.

Za skoraj vse ravni označevanja se je že izkazalo, da je mogoče boljše rezultate doseči s strojnim učenjem kot z ročno napisanimi pravili. Strojno učenje potrebuje čim boljše ročno označene učne množice, te pa potrebujemo tudi za preizkušanje kvalitete delovanja, in to tako strojnega učenja kot ročno napisanih pravil. Zato bi bilo za izboljšanje delovanja večine orodij koristno povečati količino in predvsem

raznovrstnost ročno označenih korpusov. Pri tem ni nujno, da se označijo celotna besedila, saj lahko z metodami aktivnega učenja za ročno označevanje izbiramo samo primere, ki bodo označevalniku najbolj pomagali pri izboljšanju naučenega modela. Za različne ravni označevanja bi bilo koristno tudi povečanje podpornih podatkovnih virov, predvsem leksikonov in leksikalnih baz, saj ti nudijo veliko informacij o jeziku v že prečiščeni obliki.

Izpostaviti velja tudi konceptualni model takšnega označevanja oz. povečanja podpornih virov, ki gradi na »krepostnem krogu«: z dodatnimi ročno označenimi korpusi programe naučimo boljšega označevanja, s čimer lahko pripravimo boljšo osnovo za nadaljnji krog ročnega označevanja, ta krog oz. spiralo pa lahko ponovimo večkrat.

Za točnost vseh ravni označevanja je še posebej pomemben tokenizator, saj se njegove napake prenesejo v vse naslednje stopnje označevanja; napake v tokenizaciji pa seveda neposredno onemogočajo tudi iskanje napačno tokeniziranih besed. Za slovenščino je bilo vloženih že kar nekaj naporov v izgradnjo referenčnega tokenizatorja (Krek 2011), ki je v dobršni meri tudi že implementiran v sistemu Obeliks. Seveda bi bilo možno delovanje še izboljšati; tako npr. tokenizator trenutno ne prepozna »ga.« kot okrajšave v stavkih tipa »Spoštovana ga. Micka!«. Ob tem pa se je treba zavedati, da vsaka sprememba tokenizacije glede na že obstoječe (tudi ročno) označene korpuse pomeni medsebojno neskladje virov, kar ima negativne posledice tako za označevanje kot za luščenje slovarskih podatkov iz korpusov. Take primere najdemo, če npr. skupaj uporabljamo korpuse, označene s ToTaLe, in korpuse, označene z Obeliksom. Raziskava specifično spletnega besedišča, v kateri smo iskali ključne besede slWaCa glede na referenčni korpus KRES, je pokazala, da so najbolj »ključne« prav besede, ki so s ToTaLe tokenizirane drugače kot pa z Obeliksom (Erjavec in Ljubešić 2014).

Pri oblikoskladenjskem označevanju bi bilo koristno izvesti eksperimente, s katerimi bi ugotovili, katere metode oz. kombinacije metod res dajo najboljše rezultate. Tako je bilo npr. že pokazano, da daje uporaba metaučenja, ki kombinira rezultata Amebisovega označevalnika, ki temelji na pravilih, in statističnega označevalnika TnT, boljše rezultate kot pa katerikoli od obeh samostojnih označevalnikov (Rupnik et al. 2010).

3.3 Izboljšave tehničnih vidikov orodij

Poleg točnosti označevanja posameznih orodij bi bilo koristno orodja izboljšati tudi po njihovi tehnični plati, torej pri enostavnosti namestitve in uporabe ter pri možnostih in načinih njihove integracije.

V splošnem je najbolje uporabljati odprtokodna jezikovno in od računalniške platforme neodvisna orodja, ki temeljijo na strojnem učenju, so dobro dokumentirana, se vzdržujejo na kateri od platform za kontrolo sprememb (revision control systems), kot je npr. Git, ter imajo aktivno skupino razvijalcev in uporabnikov, vključno s forumom za vprašanja, prijavo napak ali predlogi izboljšav. Primera sta npr. Moses in, do neke mere, StanfordNER. Namensko razvita orodja za slovenski jezik imajo sicer lahko prednost, da jih lažje prilagodimo specifikam svojega jezika (ali še bolj kot to: jezikovnoteoretičnemu okviru), vendar pa je vprašljivo, ali delo z njihovim vzdrževanjem in prenosljivostjo to odtehta. Taka orodja sicer v neki točki razvoja lahko dajo boljše rezultate, vendar pa je zelo verjetno, da bo razvoj strojnega učenja prinesel vse boljše rezultate. Zato je bolj smiselno trud vlagati v razvoj označenih korpusov slovenskega jezika, ki lahko služijo kot dobre učne množice, kot pa v kompleksna na pravilih temelječa orodja namenjena samo slovenskemu jeziku.

V kontekstu izdelave označenih korpusov za slovaropisje lahko zavzamemo tudi stališče, da dostopnost orodij v resnici ni pomembna, vse dokler so bila uspešno uporabljena za označitev korpusa, vendar to otežuje razširitev in tudi vzdrževanje korpusov, poleg tega pa potem ta orodja ne bodo uporabna za druge namene in označevanje drugih, s slovaropisnim projektom nepovezanih projektov oz. raziskav. Z uporabo zaprtih orodij rezultati označevanja tudi niso preverljivi oz. ponovljivi.

Naslednje vprašanje je povezljivost posameznih orodij, tako glede na njihove vhodne in izhodne formate kot tudi glede vezanosti na določene računalniške platforme. Tako je npr. že omenjena implementacija Obeliksa, sicer verjetno trenutno najboljšega dostopnega oblikoskladenjskega označevalnika za slovenščino, vezana na operacijski sistem Windows, kar ga naredi slabo kompatibilnega z okoljem Linux, ki je tradicionalno bistveno bolj opremljeno z odprtokodnimi označevalniki in drugimi orodji. Po drugi strani pa je ToTaLe vezan na Linux, kar otežuje njegovo uporabo pod operacijskim sistemom Windows. Vendar pa se v zadnjem času pojavlja vedno več platform, ki omogočajo določanje in izvajanje spletnih delotokov, kot npr. WebLicht (Hinrichs et al. 2010). V takšnih sistemih posamezni programi oz. moduli tečejo kot spletni servisi na razpršenih računalnikih, medtem ko izvajanje delotoka kot celote (npr. tokenizacija → oblikoskladenjsko označevanje → lematizacija) prevzame centralni strežnik, ki po potrebi kliče spletne servise. Mogoče je tak način izvajanja označevanja res prihodnost, vendar pa je – posebej za obdelavo velikih korpusov – trenutna rešitev še vedno predvsem v lokalnem izvajanju označevanja na v gruče povezanih računalnikih, tipično z velikimi procesorskimi in pomnilniškimi kapacitetami. Zato je pomembno, da so označevalna orodja neodvisna od operacijskega sistema oz. platforme, na kateri se izvajajo. V praksi to pomeni, da so napisana v enem od standardnih odprtokodnih programskih jezikov, kot je npr. Java ali Python.

Poleg neodvisnosti od platforme je potrebno zagotoviti specifikacije medsebojno kompatibilnih vhodnih in izhodnih formatov orodij, podobno kot je bilo npr. narejeno v sklopu nemškega sistema spletnih orodij WebLicht; več o tem piše v razdelku 4.

Zanimiv raziskovalni in razvojni izziv je torej tudi arhitektura sistemov za označevanje besedil. Skoraj vse trenutne implementacije delujejo z izbiro najboljšega kandidata (glede na model) pri vsakem koraku označevanja. Vendar pa se najboljši kandidat pri nekem koraku lahko izkaže kot neustrezen, ko dobimo na voljo več informacij, saj šele kasnejše stopnje obdelave pravilno razdvoumijo med možnimi kandidati. Tako npr. šele z upoštevanjem skladnje lahko določimo, da »ga.« ni okrajšava, temveč zaimek, ki mu sledi konec stavka, kot v že omenjenem primeru »Videl sem ga. Micka ga je tudi videla.« Novejši trendi na tem področju so sistemi, ki namesto enostavnega cevovoda označevalnikov uporabljajo Bayesove mreže (Finkel et al. 2006), v katerih vsak označevalnik ustreza eni spremenljivki sistema, s čimer je nato mogoče izvesti približno sklepanje, ki najde globalno najboljše oznake.

3.4 Predlog verige označevalnih orodij

Za označevanje korpusov slovenskega jezika za namene slovaropisja smo predvideli tiste ravni, za katere je že sedaj mogoče uporabiti obstoječa orodja, raje kot vse možne oz. potencialno koristne, ki pa bi jih bilo potrebno najprej še (skoraj) v celoti razviti. V nadaljevanju podamo predlog verige orodij za označevanje, pri čemer smo izmed predstavljenih orodij izbrali tista, ki so odprtokodna tako glede programske opreme kot tudi glede modelov slovenščine. Pri vsakem orodju navajamo tudi razmeroma enostavne predloge za njegove izboljšave.

- **Obeliks**: tokenizacija, oblikoskladenjsko označevanje in lematizacija. Sistem bi bilo koristno ponovno implementirati v enem od standardnih programskih jezikov ter ga na novo naučiti oblikoskladenjskih in lematizacijskih modelov. Pri tem bi bilo koristno dodati možnost normalizacije besed, ki bi bila sicer implementirana v ločenem modulu.
- **Moses**: normalizacija besednih oblik, pri čemer bi bil program pospremljen z več modeli normalizacije, vsaj enim za nestandardno sodobno in drugim za zgodovinsko slovenščino. Odločitev, ali neko besedilo normalizirati in s katerim modelom, bi bila lahko bodisi avtomatska glede na vsebino posameznega besedila ali pa na osnovi metapodatkov, pripisanih besedilu.
- **MSTParser**: površinskoodvisnostno skladenjsko označevanje. Za označevanje lahko uporablja obstoječi model za skladenjsko analizo, pri čemer bi bilo koristno implementirati konverzijo med shemo JOS in slovenskimi Universal Dependencies ter razčlenjevalnik naučiti še teh.

Koristno bi bilo tudi mestoma popraviti učni korpus ssj500k in ga mogoče tudi povečati z zvrstmi besedil, ki so zaenkrat slabo zastopane v korpusu, se pa predvideva, da so skladenjsko drugačne kot pa že zajete. Koristno bi bilo tudi izvesti eksperimente s kakšnim drugim, sodobnejšim skladenjskim označevalnikom, saj bi mogoče dobili s tem boljše rezultate kot z MSTParserjem.

- **StanfordNER**: določanje imenskih entitet. Učno množico (trenutno samo ssj500k) bi bilo koristno povečati in predvsem narediti bolj raznovrstno.
- Kot rečeno, za označevanje terminov trenutno ni na voljo vzdrževanega in odprto dostopnega orodja, zato bi bilo luščenje terminologije verjetno treba programirati na novo, pri čemer pa bi bilo koristno uporabiti že izdelane nize oblikoskladenjskih vzorcev, ki predstavljajo potencialne termine.

Odprto ostaja še vprašanje, kako zgoraj našteta in precej raznovrstna orodja medsebojno povezati. Za učinkovito avtomatsko označevanje velikih korpusov je najboljša rešitev instalacija in paralelizacija verige označevalnikov na visokozmogljivih strežnikih Linux oz. gručah takšnih strežnikov, pri čemer je treba pretvorbo njihovih vhodno-izhodnih formatov implementirati tako, da so medsebojno kompatibilni. Opcija tu je sicer neposredno format TEI, predviden za končni zapis korpusa, vendar je za bolj učinkovito delovanje označevalnikov predvsem primeren zapis s kazalci, vse to pa obravnavamo v nadaljevanju.

4 Zapis oznak v korpusih

Korpusi imajo lahko zelo kompleksno strukturo, tako v metapodatkih kot jezikoslovnih oznakah. V Sloveniji so se za njihov zapis v veliki meri uveljavile smernice TEI (TEI 2013), ki pokrivajo vse zgoraj obravnavane ravni označevanja kot tudi nekatere druge. Smernice so dobro vzdrževane, saj zanje skrbi mednarodni konzorcij TEI, spremlja pa jih tudi obilica orodij za izdelavo konkretnih shem XML in pretvorb iz različnih formatov in vanje, kot je npr. Word v TEI ali TEI v HTML. Elementi TEI so poslovenjeni, izoblikovala pa se je tudi večja skupina uporabnikov na področju digitalne humanistike (Erjavec et al. 2004; Ogrin et al. 2013).

Skupno priporočilom TEI je, da večino jezikoslovnih oznak zapišemo kot element XML, npr. <w> za besedo ali <name> za ime. Prednost takšnega načina zapisa je neposredna razvidnost oznak in formalna preverljivost pravilnosti zapisa, oznake in tudi besedilo pa je razmeroma lahko popravljati. Rešitev z oznakami, ki so pridružene besedilu, ima tudi več potencialnih slabosti: oznake morajo biti pravilno gnezdene (XML podpira predvsem drevesne strukture), ob večanju števila oznak postanejo elementi XML nepregledni in datoteke z vsemi vsebovanimi

oznakami zelo velike. Zato se za sisteme, v katerih se predvideva povsem avtomatsko označevanje, pogosteje uporabljajo zapisi s kazalci: vhodno besedilo se ne spreminja, oznake posameznega orodja pa kažejo na ustrezna mesta v besedilu oz. v plasti katerih drugih oznak. Tak pristop npr. uporablja že omenjeni WebLicht (Hinrichs et al. 2010) oz. njegov skupni format za zapis korpusov TCF, je pa to tudi pristop, ki ga definira standard MAF, namenjen označevanju oblikoskladnje (ISO 24611, 2012).

Čeprav je zapis s kazalci tehnično bolj enostaven in omogoča večjo fleksibilnost, pri njem dosti težje odkrijemo napake in težje povežemo posamezne ravni označevanja. Predvsem pa podatkov, na katere oznake kažejo, nikakor ne smemo spreminjati, saj s tem kazalci postanejo neveljavni. To se izkaže za problematično v primerih, ko bi hoteli ročno oz. polavtomatsko popraviti (nekatere) oznake ali samo besedilo. Zapis TEI je tako primeren predvsem za referenčne korpuse, pri katerih bi želeli imeti čim bolje preverjene oznake in zapis korpusa, za arhivske namene pa besedilo shranjeno skupaj z vsemi oznakami na čim bolj berljiv način.

Tudi večina korpusov, ki se omenjajo v tej monografiji, je zapisana v TEI, vendar je uporaba teh priporočil kompleksna, poleg tega pa so se v več kot dveh desetletjih, odkar jih v Sloveniji uporabljamo za zapis korpusov, tudi spreminjala. Poleg tega se ob dodajanju novih oznak lahko izkaže, da so bile pretekle odločitve slabo posplošljive. Zato bi bilo koristno obstoječe korpuse ne samo ponovno označiti z boljšimi orodji in modeli, temveč tudi poenotiti njihovo kodiranje, ki bi nato postalo referenčno za nove korpuse, ki bodo potrebni za izdelavo slovarja sodobnega slovenskega jezika.

5 Zaključek

V prispevku smo podali pregled ravni predvsem avtomatskega označevanja, ki bi jih bilo smiselno zapisati v korpuse, ki bodo služili bodisi kot osnova za leksikografski opis sodobnega slovenskega jezika, z njim povezanih slovarskih ali drugih virov, ali za druge namene. Izpostavili smo serijo odprtokodnih in večinoma jezikovno neodvisnih jezikoslovnih označevalnikov ter njihovih modelov za slovenski jezik, kot tudi predloge za njihove izboljšave. Na kratko je bil opisan še predlagani sistem zapisa oznak, torej format korpus.

Oznake, zapisane v velikih korpusih, so vedno dodane avtomatsko, zato se je pri uporabi takšnih korpusov treba zavedati, da orodja delajo tudi napake, te pa imajo za posledico slabše luščenje slovaropisno ali drugače zanimivih informacij. Izboljšanje točnosti teh orodij torej tudi za naprej ostaja prednostna naloga.

V
Sodobna leksikografija v teoriji in praksi

Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika

Polona Gantar, Iztok Kosem in Simon Krek

Abstract
This paper describes each stage in the compilation of a database that is to be used as a basis for an online dictionary of contemporary Slovenian and in developing Slovenian language technologies. A proposal for a procedure for archiving different versions of entries, as well as different versions of the entire database during the compilation process, is also presented. Furthermore, we describe how to include detecting lexical change (the continuous updating of headwords) and dictionary users in the process. This is a topical issue in electronic lexicography, but one that still leaves many questions unanswered.

Keywords: dictionary-making process, automatic data extraction, online dictionary, detecting lexical change, gradual dictionary compilation

Ključne besede: leksikografski proces, avtomatsko luščenje podatkov, spletni slovar, detektiranje pomenskih sprememb, postopna izdelava slovarja

1 UVOD

Izdelava slovarskih priročnikov je v digitalni dobi po pričakovanju pogojena s sodobnim načinom življenja, ki ga na področju dostopa do najrazličnejših informacij prek spleta in mobilnih naprav, kot so pametni telefoni in tablice, določajo zanesljivost, hitra in brezplačna dostopnost ter prilagodljivost vsebin, če izpostavimo samo tri najbolj odločujoče (Müller-Spitzer et al. 2011). V takih okoliščinah se leksikografi in založniki upravičeno sprašujejo, kako kompleksne slovarske opise izdelati kvalitetno, vendar hkrati v čim krajšem času in s čim manjšim finančnim vložkom ter kako jih ohranjati aktualne in zanimive za jezikovno skupnost, ki so ji v prvi vrsti namenjeni. Povsem jasno je, kot ugotavljajo poznavalci leksikografske prakse zadnjih 10, 15 let (prim. Krek 2011; Rundell 2014),[1] da tiskani slovarji, kljub temu da danes še sobivajo z elektronskimi in spletnimi, v bližnji, še manj pa daljnji prihodnosti, niso več realnost. Zato se zdi načrtovanje procesa izdelave slovarja – še posebno v situaciji, kot je slovenska, kjer na korpusu temelječi jezikovni opisi sodobne slovenščine niso na voljo, izdelava novega slovarja slovenskega knjižnega jezika (NSSKJ)[2] pa konceptualno in izvedbeno sledi tiskani logiki –, še toliko bolj pomembno in za jezikovno skupnost ključno.

V Predlogu za izdelavo slovarja sodobnega slovenskega jezika (SSSJ; Krek et al. 2013b: 52–60) je predstavljen postopek izdelave slovarja v posameznih fazah, ki omogočajo postopno objavo slovarskih informacij glede na stopnjo leksikografske obdelave in količino podatkov, ki jih imajo gesla v posameznih fazah leksikografskega opisa. Hkrati je opisan tudi postopek sprotnega posodabljanja slovarskih gesel (ibid.: 46) in določitev prioritetnega zaporedja njihove obdelave (ibid.: 45). V pričujočem prispevku želimo posamezne faze podrobneje razčleniti in se osredotočiti na načrtovanje leksikografskega procesa, ki bo kos izzivom na dolgi rok in bo učinkovito izrabljal možnosti informacijsko-komunikacijskega znanja – konkretno jezikovnotehnoloških orodij – tako v metodološkem smislu pri pridobivanju in obdelavi jezikovnih podatkov kot pri posredovanju informacij uporabnikom. Ker je proces leksikografskega opisa za spletno zasnovane na korpusu temelječe slovarje še relativno neopisan,[3] se želimo v prispevku dotakniti tudi vse pomembnejšega vključevanja jezikovne

1 Na temo prihodnosti leksikografije je bila na konferenci Electronic Lexicography in the 21st Century (eLex, Bled, 10.–12. november 2011) organizirana okrogla miza z naslovom: *Will there still be dictionaries in 2020?* Posnetek je dostopen na: http://videolectures.net/elex2011_bled/ (dostop 27. 7. 2015).

2 Slovar že nastaja na Inštitutu za slovenski jezik Frana Ramovša (ISJFR). Ker gre za koncept, ki ima na področju sodobnih jezikovnih opisov edini finančno podporo na nacionalni ravni, je problem izvedbe, ki je konceptualno vezana na tiskani format, statičnost in odsotnost jezikovnotehnološke podprtosti, dejansko problem celotne jezikovne skupnosti tako v smislu smotrnosti porabe davkoplačevalskega denarja kot v smislu rezultata, ki ne upošteva leksikografskih in jezikovnotehnoloških trendov pri jezikovnem opisu.

3 To je tudi eden od razlogov, da je bil opis in načrtovanje leksikografskega procesa pri slovarjih, zasnovanih za splet, tema ene od delavnic evropske pobude European Network of e-Lexicography (ENeL) julija 2014 v Bolzanu. Prispevki so dostopni na: http://www.elexicography.eu/working-groups/working-group-3/wg3-meetings/wg3-bolzano-meeting/ (dostop 27. 7. 2015).

skupnosti v leksikografski proces in se opredeliti do problema, ki ga prinaša sprotno objavljanje slovarskih informacij, namreč vzpostavitev postopka arhiviranja in dostopanja do posameznih različic slovarske baze.

2 FAZE V LEKSIKOGRAFSKEM PROCESU

Leksikografski proces kot detajlno načrtovan proces izdelave slovarja je ena ključnih organizacijsko-izvedbenih nalog, ki vplivajo tako na organizacijo in sestavo leksikografskega tima, kot na finančno in časovno izvedbo projekta. Kot izpostavljata Tiberius in Krek (2014), obstajajo v literaturi predvsem opisi leksikografskega procesa pri izdelavi tiskanih slovarjev (prim. Dubois 1990; Landau 1984; Zgusta 1971), kjer se načeloma izpostavljajo tri zaporedne faze, tj. faza načrtovanja, faza pisanja in faza publiciranja. Možnost izrabe računalnika (tj. strojne obdelave jezikovnih podatkov), pojav interneta in količina ter medsebojni preplet različnih tako jezikovnih kot jezikovno povezanih informacij so nujno vplivali na način izdelave in objave leksikografskih vsebin. Leksikografski proces pri izdelavi nestatičnih spletnih slovarjev, kot ugotavlja Klosa (2013: 4), tako v splošnem predvideva šest faz, ki pa ne potekajo nujno v linearnem zaporedju, ampak se med seboj lahko prekrivajo in dopolnjujejo (Klosa 2013; Tiberius in Schoonheim 2015), in sicer: pripravljalna faza, faza pridobivanja podatkov, faza računalniške priprave podatkov, faza računalniške obdelave podatkov, analitična faza in faza priprave za spletno objavo.

2.1 Opisi leksikografskega procesa v predlogih za izdelavo novega slovarja slovenskega jezika

Na področju slovenske leksikografije se je v zadnjih petih letih, dejansko pa šele od objave do pred kratkim edinega javno predstavljenega Predloga za izdelavo SSSJ (Krek et al. 2013b), začelo resneje govoriti o tem, da slovenska jezikovna skupnost potrebuje slovar sodobne slovenščine, hkrati pa tudi o tem, v kolikšni meri naj bi se leksikografska praksa pri izdelavi slovarja v digitalni dobi še zanašala na leksikografsko tradicijo (in jezikovne opise) Slovarja slovenskega knjižnega jezika (SSKJ) (Gantar 2014) ter kako tako obsežen projekt izpeljati v čim krajšem času. V kontekstu leksikografskega procesa, ki je tesno povezan z vsebino, izvedbo in medijem, za katerega je slovar zasnovan, nas bo zato zanimalo, v kolikšni meri se predstavljeni leksikografski koncepti ukvarjajo z leksikografskim procesom, koliko je ta premišljen in izvedljiv in koliko se pri tem upošteva dejstvo, da bo novi slovar namenjen predvsem bodočim uporabnikom, tj. uporabnikom, ki bodo v slovarju iskali informacije npr. čez 10, 15 let.

Preden se osredotočimo na opis posameznih faz pri izdelavi SSSJ, kot ga predvideva v ta namen oblikovan konzorcij v okviru Centra za jezikovne vire in tehnologije,[4] si poglejmo, kako je zamišljena izdelava NSSKJ, katerega izdelava naj bi trajala vsaj 20 let,[5] in še prej, kakšno vlogo ima v metodološkem smislu in z vidika vključevanja gradiva v NSSKJ Sprotni slovar slovenskega jezika, ki se spogleduje s konceptom »slovarja v izdelavi«[6] (angl. *dictionary under construction*, prim. Klosa 2013: 3), kar predstavlja v slovenski leksikografiji novost.

2.1.1 Sprotni slovar slovenskega jezika

V uvodu v Sprotni slovar slovenskega jezika lahko beremo, da gre za rastoči slovar, ki pa ima v času nastajanja in objave zgolj informativno naravo.[7] Njegova izhodiščna različica vključuje besedišče, ki je v obstoječih slovarjih še neregistrirano, potrjuje pa ga korpusno gradivo. Geslovnik se postopoma dopolnjuje z besedjem, ki so ga uporabniki iskali, a ne našli, v slovarjih na inštitutski spletni strani http://bos.zrc-sazu.si, poleg tega naj bi zajemal tudi besede, ki jih obstoječi korpusi slovenščine še ne prinašajo, raba pa je bila registrirana v drugih, zlasti elektronskih virih.[8] V zvezi s procesom in metodologijo izdelave slovarja v uvodu izvemo še, da se bo na podoben način geslovnik dopolnjeval tudi v prihodnje in da bo »novo besedje slovarju predvidoma dodajano vsakih šest mesecev«.[9] Slovar sicer omenja »izhodiščno različico«, ne predvideva pa načina arhiviranja in dokumentiranja starejših različic ter dostopa do njih, negotov pa je tudi status vključevanja v NSSKJ, ki je v uvodu opredeljen takole: »Ali bodo posamezne iztočnice dejansko prešle v normativne ali razlagalne slovarje in bodo opisane natančneje, pa bo pokazal čas.« Kljub dobrodošli novosti, ki jo obeta slovar v naslovu, lahko ugotovimo, da v metodološkem smislu, tj. z vidika postopnega dodajanja slovarskih informacij, ne prinaša v leksikografsko prakso nič novega. Kot je mogoče razbrati iz uvoda, se gesla v celoti dodajajo na novo, žal pa tudi ni mogoče preizkusiti postopka dostopanja do posameznih različic.

4 http://www.cjvt.si/projekti/ (dostop 27. 7. 2015).

5 Prim. odzive v medijih ob objavi Osnutka, npr. http://www.24ur.com/novice/slovenija/na-nov-slovar-slovenskega-knjiznega-jezika-bomo-cakali-se-leta.html (dostop 27. 7. 2015).

6 Morda ustreznejši slovenski izraz, ki označuje ta tip spletnega slovarja, je »nikoli dokončani slovar«.

7 Avtor in strokovni pregledovalci se pri izboru in opisu besedišča sklicujejo izključno na informativnost in izrecno zanikajo kakršnokoli normativnost. Zdi se torej, da bo šele premislek uredniške ekipe NSSKJ uporabnika seznanil s primernostjo, ustreznostjo oz. neustreznostjo besed. Ob povabilu uporabnikom, naj predlagajo domače ustreznike, ki jih uporabljajo ali bi jih želeli (!) uporabljati – s čimer naj bi se krepilo »zavedanje govorcev o njihovem vplivu na jezikovno rabo, preko katere lahko posledično sodelujejo pri normiranju slovenskega besedja« (Krvina 2014: 91) – pa se dejansko zastavlja vprašanje razumevanja jezikovne norme in njenega določevalca: ali o njej torej odloča skupina ljudi s pozicije moči, kot si to zamišlja ISJFR pod okriljem SAZU, ki v smislu prijaznosti in ljudskosti občasno ponudi ta občutek moči tudi izbranim jezikovnim uporabnikom, ali pa je norma – torej tisti del jezika, ki ga jezikovna skupnost želi standardizirati – neločljivo povezana s tem, kako jezikovna skupnost jezik dejansko uporablja, kar pomeni, da je izhodišče standarda v ustrezno analizirani in interpretirani jezikovni rabi celotne jezikovne skupnosti? V zvezi s tem prim. prispevek Gorjanc et al. (2015).

8 Avtorji teh virov posebej ne navajajo.

9 Kot je razvidno iz kolofona, je od leta 2014 (in tudi v času pisanja tega prispevka) na spletnem portalu Fran (http://www.fran.si/132/sssj-sprotni-slovar-slovenskega-jezika, dostop 27. 7. 2015) še vedno na voljo le različica 1.0, zadnja sprememba pa je datirana z 2. 10. 2014.

2.1.2 Osnutek za NSSKJ

Proces izdelave NSSKJ je v kontekstu pregleda leksikografskih procesov potrebno omeniti predvsem zato, ker naj bi vključeval tri pomembne elemente v t. i. predredakcijski fazi, ki se prekrivajo s predvidenimi postopki v posameznih fazah pri izdelavi SSSJ (Krek et al. 2013b), in sicer (a) avtomatsko luščenje podatkov iz korpusa (b) izdelavo orodja za prepoznavanje pomenskih in slovničnih sprememb ter (c) nadgradnjo obstoječih korpusov.

Proces izdelave NSSKJ predvideva dve fazi, t. i. predredakcijsko fazo in fazo redakcije. Predredakcijska faza vključuje oblikovanje geslovnika, na podlagi katerega bo izdelan nabor iztočnic za uvrstitev v slovar. Avtorji predvidevajo, da bodo v okviru predpriprave slovarskim podatkom v slovarski bazi samodejno dodatni podatki, pridobljeni iz korpusov, kot tudi podatki iz obstoječih slovarjev[10] in jezikovnih zbirk v lasti ISJFR.[11] Med drugim naj bi bili iz korpusov avtomatsko izluščeni zapis iztočnice, besednovrstna opredelitev, podatek o pogostosti leme in njenih posameznih oblik, skladenjski podatki s pripadajočimi kolokacijami in stavčnimi zgledi ter nekatera slovnična opozorila ter podatki o jezikovni rabi, npr. o pisanju skupaj in narazen. Na podlagi nadaljnje analize teh podatkov pa bo izdelana celostna tipologija slovarske obravnave. Ker zahteva avtomatski postopek luščenja omenjenih podatkov iz korpusa natančno razdelane odločitve glede interpretacije podatkov v razmerju korpus – leksikon besednih oblik – slovar, saj so lematizacija in oblikoskladenjski podatki prilagojeni označevanju korpusa, njihov prenos v slovar pa zato ne more biti neposreden, nas bi zanimalo, kako bo postopek avtomatizacije pri izdelavi NSSKJ dejansko izpeljan, vključno z opisom procesa avtomatizacije, saj le ta, kot se je pokazalo pri avtomatizaciji postopkov pri izdelavi dela LBS, nikakor ni trivialen. Glede na to, da omenjeni postopki na ISJFR po vsej verjetnosti še niso bili preizkušeni (oz. po vednosti avtorjev niso bili predstavljeni v strokovni literaturi, na strokovnih posvetih ali osrednjih leksikografskih konferencah po Evropi), tudi ni mogoče pričakovati rezultatov evalvacije ali natančnejše predstavitve celotnega avtomatizacijskega postopka. Če so se avtorji NSSKJ odločili, da bodo pri postopku avtomatizacije uporabili metodologijo, ki je bila uporabljena pri izdelavi LBS, je treba še poudariti, da je luščenje podatkov iz korpusa, kot je bilo izpeljano pri izdelavi LBS, sledilo zelo jasnim metodološkim izhodiščem in je bilo prilagojeno organizaciji podatkov v slovarju, ki se v marsičem razlikuje od koncepta NSSKJ.[12]

10 Glede na trditev, da bodo podatki iz obstoječih slovarjev vključeni v kar največji meri, se pod vprašaj postavlja trditev, da bo slovar »izdelan popolnoma na novo« (NSSKJ: 1, 2), posledično pa tudi, ali lahko pričakujemo dejansko opis sodobne slovenščine ali ponovno adaptacijo SSKJ, ki je nastal na povsem drugačnem gradivu, v povsem drugem času in z drugačno metodologijo.

11 Glede na tip podatkov, ki bodo pridobljeni samodejno, predvidevamo, da bo uporabljen identični postopek avtomatizacije, kot je bil uporabljen pri izdelavi Leksikalne baze za slovenščino (LBS; Kosem et al. 2013; Kosem et al. 2013a). Ustrezni citati v Osnutku za NSSKJ sicer niso navedeni.

12 Sem denimo sodi drugačna obravnava besednovrstne konverzije in homonimije nasproti večpomenskosti, obravnava iztočnice v odnosu do pomena, sistem označevanja, ki temelji na komunikacijski obvestilnosti ipd.

Na podlagi tega lahko sklepamo, da so avtorji NSSKJ zamisel o izvedbi avtomatizacije in prilagoditvi na drugače zasnovan slovar bodisi prepustili kasnejšemu času, pri čemer to vpliva na prilagajanje celotnega leksikografskega procesa, ali pa se bodo temu postopku pri dejanski izvedbi enostavno odrekli. Če si sposodimo misel iz uvoda v Sprotni slovar slovenskega jezika, lahko rečemo, da bo o vključenosti in izvedljivosti postopka avtomatizacije pri izdelavi NSSKJ dejansko pokazal šele čas.

Poleg avtomatsko izluščenih podatkov se v procesu izdelave NSSKJ predvideva tudi izdelava orodja, ki bi prepoznavalo morebitne pomenske in slovnične spremembe (NSSKJ: 78). Na ta način naj bi se po besedah avtorjev skrajšal čas izdelave slovarja, saj naj bi redaktorji dobili izbrane podatke vnaprej pripravljene za delo v leksikografskem programu, hkrati pa se v nasprotju z namenom avtomatizacije (prim. Kosem et al. 2013) predvideva, da se »pri redakcijskem delu vsi podatki preverjajo, kot da bi jih bilo treba zbrati in obdelati povsem na novo« (NSSKJ: 78).

Naslednji pomemben element, ki ga izpostavljajo avtorji NSSKJ v procesu izdelave slovarja, je razvoj orodja za avtomatsko detektiranje pomenskih in slovničnih sprememb v jeziku. Glede na to, da gre za temo, ki je v sodobni leksikografski metodologiji zelo aktualna, bi pričakovali, da gre za orodje, ki je na slovenskem gradivu že preverjeno, rezultati pa opisani v katerem od znanstvenih prispevkov. Znanje, ki ga ima na tem področju ISJFR, bi bilo namreč mogoče vsaj posredno uporabiti tudi za druge jezike, posledično pa bi se sodobna slovenska leksikografija uvrstila na pomembno mesto znotraj evropske leksikografske prakse, kjer se avtomatska detekcija pomenskih sprememb šele uveljavlja (prim. Cook et al. 2014).

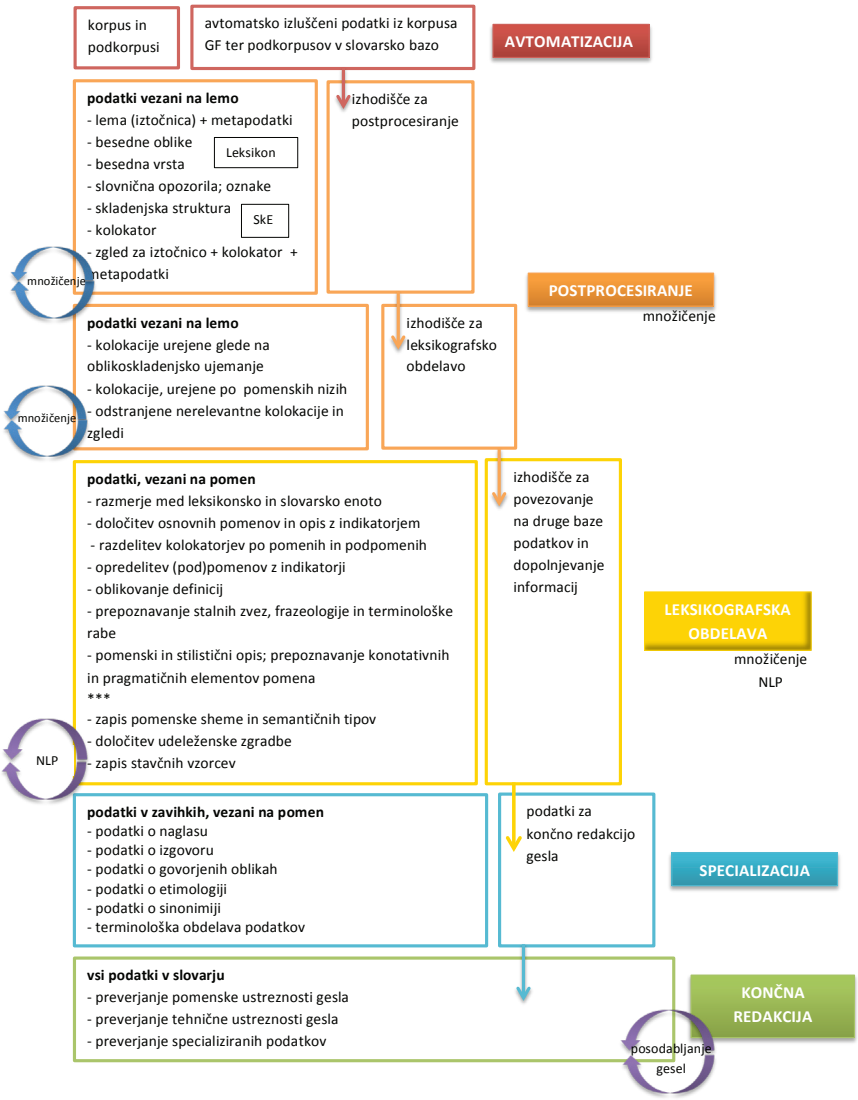
Sprotno posodabljanje korpusa je pri snovanju sodobnega slovarja, katerega uporabna vrednost ne poteče hkrati z njegovo objavo, logični postopek znotraj leksikografskega procesa. Ker pa naloga nikakor ni trivialna, bi od avtorjev osnutka za izdelavo NSSKJ pričakovali natančnejša pojasnila o tem, na kakšen način se bodo obstoječi korpusi posodabljali, do kakšne mere se bo nadgrajevala oz. spreminjala taksonomija v korpus vključenih besedil, kako bodo za nova besedila urejene avtorske pravice, kakšna bo dostopnost nadgrajenega korpusa itd. Za razliko od predvidenega postopka nadgrajevanja korpusa Gigafida pri izdelavi SSSJ, ki je opisan v pričujoči monografiji (Logar et al. 2015), namreč v leksikografskem procesu za izdelavo NSSKJ ni – razen pojasnila, da bodo »vse spremembe, ki bodo nastale med posameznimi posodobitvami slovarja, dokumentirane« ter da bo »zaradi sklicevanja na posamezne različice slovarja na voljo tudi ogled starejših različic« (NSSKJ: opombi 4 in 32), povedano nič konkretnega.

2.2 Leksikografski proces pri izdelavi SSSJ

Posamezne faze v izdelavi SSSJ so predvidene že v Predlogu (Krek et al. 2013b: 52), na tem mestu pa jih želimo razčleniti bolj podrobno, tudi v skladu z izkušnjami, ki smo jih pridobili pri nadgradnji procesa avtomatizacije, pri evalvaciji avtomatskega postopka in na podlagi izkušenj, ki jih imajo na tem področju slovarji, ki so primarno zasnovani za splet in predvidevajo postopno objavo, pri čemer so podatki na voljo uporabniku že v času nastajanja, predvideno pa je tudi sprotno posodabljanje in možnost nadgrajevanja – gre torej za slovarje, ki so na nek način nenehno v izdelavi, zaradi česar se zanje v angleščini uvaja izraz *online dictionaries under construction* (prim. Klosa 2013).

Izdelava slovarskih gesel je v okviru leksikografskega procesa pri izdelavi SSSJ predvidena v petih fazah (Slika 1) na način, ki omogoča, da so podatki uporabniku na voljo že v času izdelovanja slovarja, tj. že po prvi fazi, in ne šele po zaključku slovarskega dela. Prednosti, ki jih ima tak postopek, odtehtajo večjo zapletenost procesa, v katerem morajo biti posamezne faze zelo natančno določene, opravila leksikografov in drugih sodelavcev pa usklajena in vnaprej predvidena. Večstopenjskost izdelave gesel omogoča tudi učinkovitejšo in bolj ekonomično delitev dela. V prvi fazi tako večino dela opravi računalnik, človeško delo pa je vključeno šele v kasnejših fazah, kjer se znanje leksikografov izrablja le za specifične leksikografske postopke, ki zahtevajo izkušnje in strokovno usposobljenost. Predhodna delitev nabora iztočnic v težavnostne stopnje omogoča tudi postopno seznanjanje manj izkušenih leksikografov s postopki pomenskega členjenja in oblikovanja pomenskih razlag. Ker je za določena rutinska dela, kot je npr. odstranjevanje neustreznih in nerelevantnih podatkov ter razvrščanje kolokacij in zgledov pod ustrezne pomene, predvideno množičenje (gl. prispevek Fišer et al. 2015), je mogoče racionalizirati stroške človeških virov, predvsem pa pospešiti čas izdelave gesel.

Pri postopku izdelave slovarja v posameznih fazah kot tudi pri dostopanju do posameznih različic je zelo pomembno, da uporabnik takoj in jasno prepozna, v kateri fazi se nahaja geslo. V ta namen smo v spletni postavitvi predvideli podatek z datumom stanja gesla, ki se generira skladno s spremembami, ki potekajo v celotnem procesu geselske izdelave. Na ta način zagotovimo dvoje: prvič, uporabnik ima možnost navajanja reference na slovarsko geslo, ki velja za določeno stanje tega gesla, kar je zlasti pomembno pri citiranju podatkov v raziskovalne in izobraževalne namene, in drugič, s tem uporabniku nakažemo, kaj lahko od slovarskih podatkov pričakuje v smislu njihove količine, urejenosti in stopnje zanesljivosti.



Slika 1: Faze v leksikografskem procesu pri izdelavi SSSJ.

2.2.1 Prva faza: Avtomatizacija

Prva faza v izdelavi gesla je v celoti namenjena avtomatskemu luščenju leksikalnih podatkov iz korpusa Gigafida (Logar Berginc et al. 2012), ki predstavlja osnovno frekvenčno listo. Ta bo pred procesom luščenja dopolnjena z natančno in kompleksno statistično obravnavo podatkov iz korpusov Kres, Gos

(Verdonik in Zwitter Vitez 2011) in drugih prosto dostopnih korpusov. Poleg tega je za zajetje specializiranega besedišča predvidena izgradnja specializiranih podkorpusov in dopolnitev nekaterih že obstoječih, npr. učbeniškega podkorpusa (več o tem v Logar 2015 ter Vintar in Logar 2015), kot tudi podrobno tematsko označevanje strokovnih besedil že na ravni korpusnih metaoznak, ki se prek avtomatsko izluščenih podatkov prenesejo tudi v slovarsko bazo (prim. Gantar in Kosem 2013; Kosem 2015).

Upoštevajoč zgradbo geselskega članka, kot jo predvidevamo pri izdelavi SSSJ (prim. Klemenc et al. 2015), so avtomatsko izluščeni podatki naslednji:

- **Lema** v osnovni obliki, kot jo določa oblikoskladenjska označitev v korpusu Gigafida in leksikonu besednih oblik Sloleks (prim. Dobrovoljc et al. 2015) ter pripadajoče odvisne oblike (v samostojnem zavihku).
- Podatek o **frekvenci** leme v korpusu oz. podkorpusu.
- **Besedna vrsta** leme, kot jo določa oblikoskladenjska oznaka v korpusu in leksikonu besednih oblik Sloleks.
- Določena **slovnična opozorila**, ki se nanašajo na tipično skladenjsko ali besedilno obnašanje leme v korpusu, kot je denimo sopojavljanje z lastnimi imeni ali količinskimi izrazi, izstopajoča možnost tretjeosebne rabe glagola ali nastopanje v *se*-glagolskih ali citatnih zgradbah. Ti podatki, ki se iz korpusa pridobivajo s pomočjo kombinacije direktiv CONSTRUCTION+UNARY v orodju Sketch Engine (Kilgarriff et al. 2004), se bodisi v slovarski bazi generirajo kot opozorila leksikografom pri nadaljnji obdelavi leme v t. i. pomenski fazi oz. se lahko na ravni leme v slovar prepisujejo kot slovnične oznake, npr. *pogosto zanikano*, *pogosto v 3. os. ednine*, *pogosto z lastnim imenom* ipd. (prim. Kosem 2015).
- **Skladenjske strukture**, ki so bile predhodno registrirane na podlagi ročne analize besednih skic pri izdelavi LBS in na podlagi katerih je bila izdelana izpopolnjena različica slovničnih relacij v orodju SkE (Krek 2012). Ta različica slovnice besednih skic je namenjena izključno avtomatskemu luščenju kolokacijskih podatkov iz korpusa in ne ročnemu pregledovanju leksikografov, saj je na podlagi tako pridobljene besedne skice mogoče izluščiti bolj podrobne kolokacijske podatke, za analizo katerih bi leksikograf potreboval neprimerno več časa, zato bi zanj pomenila prej oviro kot pomoč v procesu pomenskega členjenja izhodiščne leme. Nova različica slovnice besednih skic tako vključuje tudi direktive, kot so CONSTRUCTION, COLLOC in SEPARATEPAGE, ki omogočajo luščenje vezljivostnih vzorcev pri glagolih, prepoznavanje elementov, ki na podlagi predvidenih za slovar relevantnih slovarskih enot spadajo v kategorijo t. i. skladenjskih zvez, kot so npr. zveze predlog-samostalnik-predlog: *v primerjavi z*, *v odnosu do*, *za razliko od* ipd., ter prikazovanju

relacij s tremi elementi (direktiva TRINARY) na novi spletni strani, kar omogoča uvedbo natančnejših relacij s predlogi, ki v prejšnji različici slovnice besednih skic niso bile upoštevane. Naprimer za relacijo glagol *pomesti* + predlog v tožilniku (koga-kaj_g4) dobimo samostojne stolpce s kolokatorji za različne predloge, ki se vežejo s tožilnikom: pomesti pod [preprogo, predpražnik, tepih], *pomesti na [smetišnico, smetišče, kup, tla], pomesti v [koš, kot]* ipd.

- **Kolokatorje**, ki se pojavljajo v posamezni skladenjski strukturi ob obravnavani lemi in tvorijo potencialne kolokacije pa tudi skladenjske in stalne zveze. Prepoznavanje zadnjih kot tudi uvrstitev v določen pomen je naloga leksikografa v pomenski fazi leksikografskega procesa.
- **Korpusne zglede**, ki vsebujejo lemo in kolokator v določeni skladenjski strukturi, za kar smo uporabili za slovenščino prilagojeno in v dveh korakih izpopolnjeno funkcijo orodja GDEX za izločanje čim bolj optimalnih korpusnih zgledov (Kosem et al. 2013), ki predstavljajo kandidate za vključitev v slovar (z možnimi prilagoditvami) in so kot taki torej potencialni slovarski zgledi.

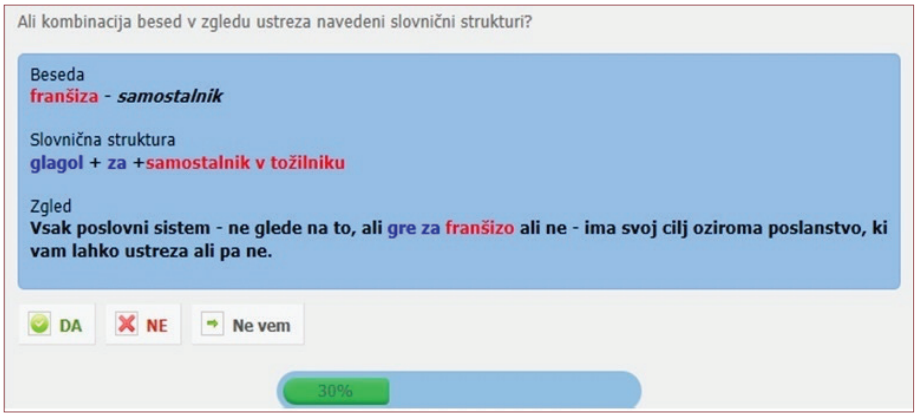
Postopek avtomatizacije, kjer smo s pomočjo v ta namen posebej prilagojene API skripte, ki vsebuje opise vseh relevantnih slovničnih relacij, izluščili zgoraj navedene podatke in jih avtomatsko prenesli v slovarsko bazo v programu iLex (Erlandsen 2004), kjer so bili pripravljeni za nadaljnjo obdelavo, je bil za slovenščino že preizkušen pri izdelavi LBS (Kosem et al. 2013; 2013a) in ovrednoten z vidika prekrivnosti izluščenih podatkov glede na ročno izdelavo gesel, v postopku izdelave kolokacijskega slovarja za slovenščino pa še nadgrajen in izboljšan. V nadgradnjo sodi predvsem avtomatsko odstranjevanje kolokatorjev, ki ponudijo same enake zglede, in postavitev leme in/ali kolokatorja pri izpisu v slovarsko bazo v ustrezen sklon, spol in število. Poleg tega smo ob predhodno natančno določenih parametrih za luščenje kolokatorjev, ki smo jih izdelali ločeno za različne frekvenčne skupine in posamezne besedne vrste (za podrobnosti gl. Kosem et al. 2013), v novem poskusu kolokatorje izluščili na podlagi združenega podatka o jakosti (angl. *salience*) in frekvenci kolokatorja, ki omogoča primerjavo z ročnim izborom kolokatorjev na podlagi besednih skic in v končni fazi omogoča izbor kolokatorjev, ki so za določeno lemo najrelevantnejši.

2.2.2 Druga faza: Postprocesiranje in odstranjevanje napak

Druga faza v procesu izdelave slovarja je namenjena (a) postprocesiranju, ki vključuje tudi čim bolj avtomatsko odstranjevanje napak oz. nerelevantnih avtomatsko izluščenih podatkov z možnostjo uporabe množičenja, (b) dodajanju metaoznak, ki omogočajo povezovanje podatkov znotraj slovarske baze in združevanje

z drugimi slovarskimi bazami (npr. izhodiščno slovarsko bazo in bazo za izdelavo kolokacijskega slovarja, slovarja sinonimov ipd.) ter (c) čiščenju nerelevantnih podatkov, ki so pri avtomatskem luščenju običajno posledica napačne lematizacije ali korpusnega šuma. V postopku postprocesiranja je avtomatsko izluščene podatke mogoče dodatno urediti, npr. postaviti izluščene kolokatorje v ustrezen spol in sklon glede na lemo, kot ga zahteva podstavna skladenjska struktura, ter vzpostaviti kolokacijske nize, v katerih so kolokatorji znotraj enega niza pomensko povezani. V postopku postprocesiranja je za namene združevanja različnih podatkovnih baz potrebno posebej označiti posamezne elemente znotraj kolokacij, npr. predloge, veznike, prosti morfem *se/si* pri glagolu ipd. ter odstraniti napačno zapisane strukture, ki so se npr. pri združevanju avtomatsko in ročno izdelanih gesel pokazale kot posledica ročnega zapisovanja. Dodajanje metaoznak posameznim elementom geselske zgradbe, npr. identifikacijska oznaka iz leksikona besednih oblik Sloleks pri lemi in kolokatorju, ter dodajanje opozoril o statusu leme, npr. o njeni avtomatski ali ročni izdelavi, je namenjena predvsem leksikografom in združevanju podatkov iz različnih baz.

Za odstranjevanje nerelevantnih kolokacij, ki se zaradi lematizacijskih napak in korpusnega šuma pojavijo pri avtomatskem luščenju podatkov iz korpusa, smo v tej fazi predvideli tudi možnost uporabe množičenja, pri katerem v posebej za to oblikovani nalogi uporabnike sprašujemo, ali uporabljena kombinacija v avtomatsko izluščenem zgledu ustreza predvideni skladenjski strukturi, kot prikazuje Slika 2:



Slika 2: Naloga v procesu množičenja za odstranjevanje avtomatsko izluščenih nerelevantnih kolokacij in pripadajočih zgledov iz LBS

Proces čiščenja avtomatsko izluščenih podatkov iz korpusa, za katerega smo uporabili orodje SlowCrowd[13] (Tavčar et al. 2012), ki se učinkovito izrablja tudi za

13 http://nl.ijs.si/slowcrowd/index.php?project=slowcrowdmain (dostop 27. 7. 2015).

čiščenje slovenske različice wordneta SloWNet (Fišer 2009), je bil preizkušen pri izdelavi LBS (Kosem et al. 2013a), prvi rezultati pa so pokazali, da je uporaba množičenja pri čiščenju podatkov zanesljiva in da lahko občutno skrajša čas v tej fazi leksikografskega procesa.

2.2.3 Tretja faza: Leksikografska obdelava podatkov

V naslednji fazi, ki je namenjena leksikografski obdelavi podatkov, zaradi česar je strokovno in organizacijsko najbolj zahtevna in hkrati časovno najobsežnejša, so naloge osredotočene na analizo podatkov z vidika pomena, tj. pomenske členitve in pomenskega opisa, posledično pa zadevajo tudi prepoznavanje slovničnih in skladenjskih, normativnih ter stilističnih lastnosti besed oz. njihovih pomenov.

V tej fazi ima leksikograf na voljo avtomatsko izluščene in izčiščene podatke za posamezno lemo, ki ji je avtomatsko pripisana besednovrstna oznaka, ki ustreza morfosintaktični oznaki v leksikonu, zato je njegova prva naloga prepoznavanje simetričnosti med t. i. leksikonsko in slovarsko enoto. Naloga nikakor ni trivialna, učinkovitost in enotnost leksikografov pri odločanju v posameznih primerih pa je povezana z natančnimi navodili, ki vključujejo vse možne situacije in predpostavljajo enotne rešitve pri obravnavi homonimije ter besednovrstne konverzije – tj. v skladu z dogovorjenimi merili slovarskega koncepta (prim. Gantar 2015; K. Dobrovoljc 2015). Konkretno je pri izdelavi SSSJ razmerje med leksikonsko in slovarsko iztočnico privzeto simetrično, morebitne posebnosti v slovarski bazi pa leksikograf označuje s posebnimi vnaprej izdelanimi strojno berljivimi restriktorji.

Osnovna naloga, ki jo v tej fazi opravi leksikograf, je pomenska razčlenitev leme in oblikovanje pomenskih razlag za prednostno določena gesla. Glede na zgradbo gesla, ki jo povzemamo po LBS, leksikograf izdela pomenski meni s pomočjo pomenskih indikatorjev, v katerem je prikazana pomenska zgradba gesla in razmerja med pomeni in podpomeni, ter t. i. pomensko shemo s prikazom tipičnega vezljivostnega vzorca za posamezni (pod)pomen pri glagolskih ter nekaterih samostalniških in pridevniških iztočnicah. Na tej stopnji je pomembno tudi dodajanje podatkov, ki so v slovarski bazi namenjeni računalniški obdelavi jezika, kamor sodi (z uporabo sofisticiranih avtomatskih postopkov) luščenje stavčnih vzorcev, prepoznavanje semantičnih tipov po vzoru Corpus Pattern Analysis (Hanks 2004; Hanks in Pustejovsky 2005) in pripisovanje udeleženskih vlog (angl. *semantic role labeling*).

V tej fazi leksikograf identificira tudi stalne besedne zveze, med katerimi posebej opozori na terminološke, ki potrebujejo obravnavo z vidika stroke, na katero se nanašajo, ter registrira in pomensko opiše frazeološke enote.

Leksikografsko delo je v tej fazi organizirano glede na stopnjo težavnosti gesla in vnaprej pripravljenih predlog za določene pomenske skupine gesel. Za učinkovito organizacijo dela je pomembna razdelitev nalog med (a) izkušene leksikografe, ki opravijo osnovno pomensko razčlenitev in izdelajo pomenske razlage, prepoznajo kompleksnejše slovnične in skladenjske vzorce, stilistične ter pragmatične posebnosti rabe ipd., med (b) leksikografe, ki so specializirani za redakcijo frazeoloških enot, opis slovničnih in skladenjskih lastnosti posameznega (pod)pomena, normativnih podatkov, ter (c) med leksikografe, ki se z leksikografskimi nalogami šele seznanjajo in opravljajo manj zahtevna leksikografska opravila, kot je preverjanje ustrezne razvrstitve kolokacij pod posamezne pomene (na podlagi rezultatov množičenja) in skladenjske strukture, urejanje pomenskih kolokacijskih nizov, prepoznavanje besedilnega okolja pri stalnih zvezah in frazeoloških enotah.

Za del nalog, ki predstavljajo rutinska leksikografska opravila in ne zahtevajo poglobljenega leksikografskega znanja, je predvidena naloga v okviru množičenja, kjer uporabniki razvrščajo avtomatsko izluščene zglede, ki vsebujejo kolokacijo, ki ustreza določeni skladenjski strukturi, pod ustrezni vnaprej določeni (pod)pomen.[14] S to nalogo želimo poleg razvrščanja kolokatorjev v že pomensko razčlenjeno geslo dobiti tudi povratno informacijo o ustrezni pomenski razčlenitvi ter detektirati pomenske opise, ki znotraj širše jezikovne skupnosti nimajo zadostne potrditve.

Po končani tretji fazi, v kateri imajo uporabniki na voljo že večino slovarsko relevantnih informacij, vezanih na pomen, je slovarsko geslo pripravljeno za dodajanje informacij, ki jih v spletno zasnovanem slovarju predvidevamo v drugih zavihkih/rubrikah, čemur je namenjena naslednja faza v leksikografskem procesu.

2.2.4 Četrta faza: Dodajanje specializiranih jezikovnih podatkov

Delo v četrti fazi postopno izdelanega slovarskega gesla je namenjeno dodajanju t. i. zunajbaznih podatkov in dopolnjevanju na spletnem portalu že prikazanih podatkov, kjer se predvideva specializirano znanje jezikoslovcev in strokovnjakov drugih področij, zlasti terminologov pa tudi jezikoslovcev, zadolženih za standardizacijo in reševanje normativnih vprašanj (gl. Popič et al. 2015). Podatki, ki se posameznim delom slovarskega gesla dodajajo v tej fazi, so naslednji:

- podatki o **naglasu** v povezavi z leksikonom besednih oblik in v skladu z že omenjenim prepoznavanjem simetričnosti leksikonske in slovarske

14 Naloga je podrobneje predstavljena v prispevku Fišer et al. (2015).

enote, kamor sodi tudi prepoznavanje prekrivnosti oz. neprekrivnosti naglasne in spregatvene paradigme kot enega izmed meril za obravnavo homonimije v odnosu do večpomenskosti;

- podatki o **izgovoru** na podlagi korpusa Gos in na podlagi za to določenih vnaprej izdelanih parametrov, kamor sodi označitev naglasnega mesta, izgovor odvisnih oblik in zapis izgovora, ki iz zapisa iztočnice ni predvidljiv (Jurgec 2015);
- podatki o **govorjenih oblikah** in posebnostih posameznih oblik na ravni pomena, kot jih je mogoče pridobiti iz govornega korpusa Gos (Verdonik 2015);
- podatki o **etimologiji,** natančneje o izvoru besede in njenih sorodnih oblikah v različnih jezikih ter podatek o starinskih oblikah ali zapisih besede v slovenskem jeziku glede na časovno umeščenost besedila, v katerem se je konkretna oblika pojavila;
- podatki o **sinonimiji** na podlagi analize kolokatorjev v funkciji Sketch Difference v orodju Sketch Engine in z vključitvijo podatkov iz slovenske različice wordneta SloWNet ter
- **terminološka obdelava podatkov.** V ta namen bo organizirana mreža strokovnjakov za posamezna strokovna področja in vzpostavljena spletna platforma, ki bo omogočala spremljanje in usklajevanje dela.

2.2.5 Peta faza: Končna redakcija gesla

Zadnja faza v leksikografskem procesu pri izdelavi SSSJ je namenjena končni redakciji celotnega gesla in medsebojnemu usklajevanju podatkov, ki jih vključujejo posamezni slovarski zavihki. Leksikografova naloga je ugotavljanje konsistentnosti podatkov glede na leksikografski koncept, strukturo gesla, pa tudi ugotavljanje skladnosti podatkov z izpričanim realnim jezikovnim stanjem. Leksikograf lahko zato pred zaključno objavo posamezno geslo uredi, dopolni ali pa vrne v katero od predhodnih faz, če npr. ugotovi nedoslednosti v pomenski členitvi besede, pri opisu stalnih zvez in frazeoloških enot ali pomanjkljive podatke v segmentu terminološke obdelave.

Pred zaključkom te faze je opravljen tudi postopek avtomatskega detektiranja pomenskih sprememb, ki vrnejo obravnavo besede v fazo pomenskega členjenja, prepoznavanja večbesednih leksikalnih enot, izrabe moči množic in ponovne končne redakcije. Čeprav torej predstavlja peta faza zaključen leksikografski proces, se slovarski opis na tej točki ponovno vrača v avtomatski izvoz podatkov, ki jih v zvezi s posameznimi pomeni narekuje leksikalni razvoj slovenščine, izpričan v nadgradnji korpusnih virov.

3 SPROTNO POSODABLJANJE SLOVARSKE BAZE

V celotnem postopku izdelave gesel in njihove predstavitve uporabnikom ima zelo pomembno vlogo slovarska baza, ki po eni strani predstavlja vir vseh slovarskih informacij, po drugi pa tudi arhiv vseh sprejetih odločitev v posameznih fazah leksikografskega procesa. Ker so faze izdelave slovarja jasno razmejene, je treba z vidika slovarske baze zagotoviti, da je v slovarskem orodju vzpostavljen delotok, ki lahko redaktorjem in leksikografom v vsakem trenutku postreže z informacijo, v kateri fazi je posamezno geslo. Dodatno raven kompleksnosti pri načrtovanju slovarske baze prinašata dve povezani odločitvi: sprotno posodabljanje slovarja in predvidena možnost, da so gesla na voljo uporabnikom po vsaki dokončani fazi.

Ko govorimo o sprotnem posodabljanju, imamo v mislih posodabljanje obstoječih gesel v slovarski bazi, ki so že šla skozi vse faze leksikografskega procesa, deloma pa tudi izdelavo povsem novih gesel, sploh tistih, ki so bila obravnavana prednostno (npr. neologizmi). Slednja morajo biti namreč v bazi opremljena s posebnim opozorilom o svoji pomembnosti, ki jih ločuje od drugih gesel, tudi zato, da se lahko pri posodobitvi slovarja uporabnike nanje opozori. Podobno velja tudi za posodabljanje obstoječih že dokončanih gesel, kjer gre lahko bodisi za dodajanje novih podatkov (npr. pomenov, kolokacij, frazeoloških enot) na podlagi analize novega gradiva (npr. spremljevalnega korpusa ali dolgoročno gledano nove verzije referenčnega korpusa) bodisi za popravljanje obstoječih (npr. popravljanje odkritih napak), vendar pa je tu za razliko od prednostnih gesel pomembnejši časovni vidik, torej kdaj so bile nove informacije dodane (in na podlagi katerega vira).

Pri sprotnem posodabljanju je treba posebej obravnavati primere, ko v slovarsko bazo dodajamo nove podatke, ki nadomestijo stare samo na ravni prikazovanja uporabnikom. Tak primer so naprimer slovarski zgledi, kjer se lahko na neki točki odločimo za zamenjavo obstoječih slovarskih zgledov z novimi (prim. Klein in Geyken 2010; Lemnitzer et al. 2015). Zaradi tega je treba v slovarsko bazo pri zgledih (in drugih mikrostrukturnih slovarskih elementih) zapisati opozorilo o tem, ali jih v slovarju prikazujemo ali ne. Na ta način v bazi vedno ohranjamo tudi vse predhodne zglede, uporabnikom pa so v določeni fazi vidni le tisti, ki so glede na vsebino gesla najbolj aktualni.

Možnost, da so gesla na voljo uporabnikom po posamezni fazi leksikografskega procesa, sama na sebi v bazi ne zahteva dodatnih informacij, razen tistih, ki se nanašajo na delotok; dobro je le iz takšnega postopka izločiti obstoječa slovarska gesla, ki jih posodabljamo z novimi informacijami, saj bi kombinacija leksikografsko pregledanih in nepregledanih podatkov v geslu lahko zmotila uporabnike. Sicer so za objavo po fazah precej bolj relevantne vizualizacijske

rešitve v slovarju, ki pa vseeno potrebujejo tudi določeno informacijo (npr. datum objave in različica).

Ključno je torej vzpostaviti postopek, ki leksikografom prikaže jasno sliko o tem, v katerih fazi izdelave se geslo nahaja, kdaj je bilo dodano v slovar, kdaj so bili geslu dodani novi podatki (kdaj je nastala nova različica[15]). Po našem mnenju lahko tak postopek zagotovi učinkovito in pregledno leksikografsko delo in omogoči jasno in razumljivo predstavljanje slovarskih podatkov uporabnikom.

4 OBRAVNAVANJE POSAMEZNIH RAZLIČIC IN VIZUALIZACIJA PODATKOV

V tem delu se posvetimo trem vprašanjem, relevantnim za posodabljanje slovarja s predlaganim leksikografskim procesom: Kako pogosto posodabljati slovar? Kako jasno ločiti nedokončana gesla oz. njihove informacije od končanih? Kako obravnavati posamezne različice gesel in slovarja kot celote?

Tuje prakse pri pogostosti posodabljanja spletnega slovarja kažejo dva pristopa: redni nekajmesečni intervali ali sproti, čim nastanejo nova gesla. Prvi pristop uporabljajo slovarji, kot je npr. Oxford English Dictionary (OED),[16] kjer slovar posodobijo vsake štiri mesece in imajo posebno stran,[17] namenjeno objavam v zvezi s posodobitvami. Podoben način posodabljanja uporablja tudi Macmillan English Dictionary (MED),[18] ki pa ne podaja ločenih informacij o tem, kdaj je posodobitev opravljena,[19] ampak uporabnike opozori na izbrane nove besede v rubriki New Words na naslovni strani.

Drugi pristop, torej takojšnjo objavo dokončanih gesel, zasledimo v Velikem slovarju poljskega jezika[20] (Żmigrodzki 2014) in v Slovarju sodobnega nizozemskega jezika[21] (Tiberius in Schoonheim 2015). Poudariti velja, da gre pri omenjenih slovarjih za projekta, pri katerih se slovarja delata na novo in sta torej sproti nastajajoča slovarja v pravem pomenu besede. Tako je motivacija za čim prejšnjo objavo leksikografskih vsebin zaradi uporabnikov, pa najbrž tudi financerjev, toliko

15 Pomembno je ločiti med različicami gesel, ki nastanejo ob večjih spremembah (npr. po končanju posamezne faze ali pri posodobitvi obstoječega slovarskega gesla z novimi podatki), in med različicami, ki jih sproti beleži program za izdelavo slovarjev. Ta namreč beleži vse sprotne spremembe, ki jih pri izdelavi gesla vnaša leksikograf, in tako omogoča primerjavo dveh različic, vpogled v izbrisane podatke, njihovo obnovitev ipd. Zato je dobro dosledno uporabljati terminologijo, ki jasno ločuje med tema procesoma.

16 http://www.oed.com/ (dostop 27. 7. 2015).

17 http://public.oed.com/the-oed-today/recent-updates-to-the-oed/ (dostop 27. 7. 2015).

18 http://www.macmillandictionary.com/ (dostop 27. 7. 2015).

19 Na spletni strani s pogostimi vprašanji in odgovori: (http://www.macmillandictionary.com/faq.html (dostop 27. 7. 2015). je podana informacija, da slovar posodobijo večkrat na leto.

20 Wielki słownik języka polskiego: http://www.wsjp.pl/ (dostop 27. 7. 2015).

21 Algemeen Nederlands Woordenboek: http://anw.inl.nl/show?page=search1 (dostop 27. 7. 2015).

večja kot pri slovarjih, ki večinoma samo dodajajo nove besede ali posodabljajo obstoječo vsebino. S tega vidika je predlog v konceptu NSSKJ, ki predvideva posodabljanje slovarja enkrat letno (NSSKJ: 3), premalo ambiciozen in premalo uporabniško naravnan – pričakovali bi namreč, da bi slovarskim uporabnikom, za katere vemo, da že več kot dvajset let čakajo na nov opis slovenskega jezika, čim hitreje ponudili rezultate leksikografskega dela.

Logiki sprotnega objavljanja sledi tudi leksikografski proces pri predlaganem SSSJ, v katerem predvidevamo sprotno objavo gesel v vseh fazah izdelave. V slovenskem prostoru sicer že obstajajo slovarji, ki uporabljajo podobno prakso, npr. iSlovar,[22] ki ločuje štiri stopnje dokončanosti gesla: »predlog« (predlagal urednik ali uporabnik), »pregledano« (pregledal urednik), »strokovno pregledano« (pregledala in uredila strokovna skupina) in »urejeno« (pregledala slovaropisna skupina; gre za končno redakcijo). Opozoriti je treba, da to ne pomeni, da se gesla dodajajo vsak dan ali celo vsako uro, ampak gre za paketne posodobitve (torej več gesel hkrati) v precej pogostih intervalih, ki zagotavljajo boljšo preglednost tako za leksikografe kot za uporabnike.

Z opozorili o dokončanosti gesla tudi poskrbimo za ločevanje nedokončanih gesel od dokončanih. V Predlogu za izdelavo SSSJ (Krek et al. 2013b: 52–60) je predvideno, da se stanje gesla prikaže z barvnimi pikami (od rdeče do zelene), hkrati pa se navede tudi datum zadnje posodobitve gesla (Krek et al. 2013b: 27). Datum je razločevalni element tudi v primeru, da geslo ostane v isti fazi (npr. pri posodobitvi dokončanega slovarskega gesla). Končna oblikovalska rešitev bo mogoče drugačna od prikazane v Predlogu, a bo v vsakem primeru morala vključevati vsaj ti dve informaciji. Poleg tega bo, podobno kot pri OED, na posebni (pod)strani slovarja dokumentirana vsaka posodobitev, npr. s seznami iztočnic in njihovim statusom, izpostavljene bodo tudi morebitne sistematične spremembe.

Ena izmed odločitev glede sprotnega objavljanja se nanaša tudi na to, kaj po spremembah narediti s prejšnjimi različicami gesel. To je sicer manj problematično v primeru objavljanja posameznih faz, saj je za uporabnika v vsakem primeru najbolj relevantna različica gesla, ki vsebuje največjo količino informacij. Na simpoziju o angleškem slovarju OED leta 2014 je bila prav to ena od perečih tem, saj so se nekateri udeleženci pritožili, da po posodobitvi gesel nimajo več vpogleda v prejšnje različice. Njihov argument je bil, da npr. s posodobitvijo razlag izgubimo informacijo o tem, kako so na določen pomen oz. rabo besede v jezikovni skupnosti gledali v obdobju izdelave prvotne različice gesla. Tak argument je seveda povsem legitimen, vendar pa velja spomniti, da gre v tem primeru za historični slovar, pri katerem je diahroni pogled na rabo jezika ključnega pomena.

22 http://www.islovar.org (dostop 27. 7. 2015).

Pri izdelavi SSSJ nameravamo o možnosti primerjanja različnih prejšnjih in zadnje različice slovarja razmisliti skladno z uporabniškimi raziskavami, kjer bi se zdelo smiselno o taki možnosti razmisliti, če bi uporabniki izrazili to potrebo.

Že sedaj predviden način dostopa do starejših različic gesel bo za raziskovalce, pa tudi za namene strojne obdelave jezika, omogočen z rednim objavljanjem novih različic prosto dostopne slovarske baze, ki bodo dejansko usklajena s posodobitvami spletne različice slovarja, razen v primerih, ko bo prišlo do sprememb v podatkih, ki so relevantni samo za slovarsko bazo. Pomemben del takšnih objav bo podrobna dokumentacija, kjer bodo opisane ne le spremembe na ravni slovarskih gesel, ampak tudi vse vsebinske in tehnične spremembe v slovarski bazi, npr. novi tipi skritih oznak, spremembe na ravni DTD (angl. *Document Type Definition*) itd.

5 ZAKLJUČEK

Leksikografski proces pri izdelavi slovarja, ki predvideva sprotno objavljanje gesel v posamezni fazi, njihovo posodabljanje in možnost dostopa do posameznih različic izdelanega gesla, je kompleksen postopek, ki zahteva vnaprej predvideno in natančno izdelano strategijo, ki vpliva na dejansko izvedljivost celotnega postopka in organizacijo leksikografskega dela. Leksikografski proces, kot ga predlagamo v prispevku, temelji v izhodišču na avtomatskem luščenju osnovnih leksikalnih podatkov, ki se v naslednjih fazah dopolnjujejo, hkrati pa se izločajo podatki, ki so bodisi napačni ali pa za uporabnika nerelevantni. Pri tem je pomembno razlikovati med slovarsko bazo in podatki v njej, ki so namenjeni tekoči izvedbi sprotnega posodabljanja in uporabniku niso vidni, in podatki, ki jih ima uporabnik na voljo v posameznih različicah. Za izvedbo predlaganega procesa je posebej pomembno, da je delotok natančno razdelan in opisan, kar leksikografom in redaktorjem omogoča kontinuirano in konsistentno izvedbo, uporabniku pa je ves čas pred očmi informacija, v kateri fazi se geslo nahaja, ter posledično, kateri podatki so mu v določeni fazi na voljo, hkrati pa tudi, kako dostopati do posameznih različic, če ga to zanima za raziskovalne namene, za nadaljnje strojno procesiranje ali izrabo podatkov v pedagoške namene. V zvezi z leksikografskim procesom, ki ga predlagamo, je pomembno tudi poudariti, da je izdelan izključno za objavo na spletu in za natančno predvideno geselsko zgradbo, notranjo organizacijo ter vrsto leksikalnogramatičnih podatkov, ki se dopolnjujejo tako znotraj geselske zgradbe (v zavihku pomen, npr. pomenski meni, kolokacije, skladenjske strukture, stavčni vzorci, zgledi pri posameznih leksikalnih enotah, tj. pomenih, stalnih zvezah in frazeoloških enotah) ter s podatki v drugih zavihkih na spletni strani, tj. s podatki o besednih oblikah, govoru, normi, sinonimiji itd.

Metamorfoze definicije v francoskem slovaropisju

Gregor Perko

Abstract
This article deals with the various forms and methods of semantic description (lexicographic definition) in modern metalexicography and French monolingual lexical production. Although the title of the article evokes a diachronic perspective, the main focus is on contemporary lexicographic production. The first part of the article attempts to define the concept of lexicographic definition. The second part analyses the relationship between lexicography and modern linguistics and its implications for semantic lexicographic description. The splitting of the concept of lexicography into *lexicographie* and *dictionnairique* opens up new possibilities for studying the definition of the term (user's perspective, dictionary functions). This section is particularly interested in two types of dictionaries: French learner's dictionaries and dictionaries designed for systematic vocabulary acquisition.

Keywords: monolingual lexicography, French lexicography, metalexicography, semantics, definition

Ključne besede: enojezično slovaropisje, francosko slovaropisje, slovaroslovje, semantika, definicija

1 UVOD

V slovenskem prostoru je bilo po strukturalnem pristopu, kakršen je bil pri definicijah izpeljan v Slovarju slovenskega knjižnega jezika (SSKJ; prim. Vidovič Muha 2000), do sedaj vprašanje slovarskih definicij predstavljeno skoraj izključno na način, kot se je uveljavil anglosaškem prostoru (Krek 2004; Gantar in Krek 2009). V svojem članku se bom posvetil različnim oblikam in možnostim pomenskega opisa, kot jih najdemo v francoski enojezični slovarski produkciji in sodobnem slovaroslovju. Čeprav naslov članka zveni diahrono-razvojno, je ključni poudarek na sodobni produkciji. V prvem delu opredelim pojem *slovarske definicije*, v drugem pa podrobneje analiziram navezavo slovaropisja na sodobne jezikoslovne pristope in posledice za pomenski opis. S cepitvijo pojma slovaropisja na slovaropisje v ožjem pomenu (fr. *lexicographie*) in slovarstvo (fr. *dictionnairique*), ki ga je v francosko slovaroslovje[1] vpeljal Quemada (1987), se odprejo nove možnosti za analizo definicije s pomembnejšim upoštevanjem uporabniškega vidika in funkcije slovarja. V tem razdelku se posebej zanimam za dva tipa slovarjev, t. i. šolske slovarje in pa slovarje, ki so namenjeni sistematičnemu usvajanju oz. učenju besedišča, s čimer se navezujem na razprave o različnih uporabnikih in njihovih potrebah v tej monografiji (Arhar Holdt 2015; Arhar Holdt et al. 2015; Čibej et al. 2015; Mikolič 2015; Rozman et al. 2015).

2 ZAKAJ DEFINICIJA?

Številni strokovni izrazi imajo znotraj posameznih strok dolgo zgodovino in njihov razvoj je pogosto nemogoče jasno začrtati. Eden takšnih je tudi *definicija*. V slovaropisju se je termin, ki označuje osrednjo informacijo o pomenu geselskega ali podgeselskega leksema, do sredine 20. stoletja uporabljal, ne da bi bil natančno definiran in opredeljen.

Termin *definicija* ima svoj izvor v logiki in označuje trodelni izrek (*definiendum* + *definor* + *definiens*), katerega namen je določiti obseg (ekstenzijo) pojma. Pojem definicije, kot ga pojmuje tudi sodobno slovaropisje, se navezuje na Aristotelovo tradicijo, ki jo povzema znani sholastični izrek: »*Definitio fit per genus proximum et differentiam specificam*«. V sodobni literaturi je takšen tip definicije poznan tudi kot analitična definicija. Pojem definiramo torej tako, da podamo najbližji rod in specifično razliko, ki ga razlikuje od ostalih pojmov, ki pripadajo istemu rodu. Takšna definicija mora zadostiti vrsti zahtev: biti mora logično pravilna, zajeti mora naravo oz. bistva pojma, ki ga definira, ne sme biti krožna, ne sme

1 Teoretsko in metodološko koristna dihotomija se izven francosko govorečega prostora ni uveljavila. O tem posredno priča tudi odločitev H. Béjointa, ki v svoji angleško pisani monografiji (Béjoint 2000: 89) dihotomijo sicer na kratko obravnava, vendar terminov ne prevaja v angleščino.

biti negativna, pojmi, ki služijo za razlago, morajo biti dovolj jasni, enostavni in enoznačni (Wiegand 1992; Uršič in Markič 1997).

Že prve sistematične slovaroslovne študije (Marcus 1970; Rey-Debove 1971; Henne 1972; Sinclair 1987; Wiegand 1989; 1992; Rey 1990) so upravičenost uporabe termina *definicija* postavile pod vprašaj. Naj na kratko povzamem. Slovarska definicija ni trodelna, ampak zajema sam *definiens*, *definiendum* (geselski leksem) ni del mikrostrukture in torej tudi ne definicije kot take, tudi *definor*, vez, v definiciji ni prisoten, je neizražen in ostaja impliciten (*je, pomeni, označuje* …). Krožnosti se nikdar ne bo mogoče izogniti v celoti, četudi je seveda moč preprečiti »neposredno« krožnost (definirati A z B in B z A). Besedišče, ki je uporabljeno v geslovniku slovarja, namreč vedno nastopa tudi v definicijah. V vseh slovarjih so dokaj pogoste tudi t. i. negativne definicije, kjer nastopa protipomenka, kar je razumljivo in smiselno, saj je protipomenskost eno od kognitivno najbolj dostopnih medleksemskih razmerij. Poglejmo preprost primer iz Nouveau Petit Robert (NPR), ki pridevnik *nesposoben* definira kot »kdor ni sposoben« in »komur manjka sposobnosti«:

> **INAPTE** /…/ qui n'est pas apte, qui manque d'aptitude.

V definicijah nujno nastopajo leksemi, ki so večpomenski, kar preprečuje popolno enoznačnost. Drugo vprašanje je, koliko je zavoljo tega definicija »nejasna«. Poglejmo primer iz NPR:

> **HAUTEUR** ['otœʀ] **nom féminin ÉTYM.** XII^{e} ◊ de *haut* Famille étymologique **haut I DIMENSION, POSITION VERTICALE A. LA HAUTEUR 1** Dimension dans le sens vertical, de la base au sommet

Vse polnopomenske besede v zgornji definiciji so v istem slovarju obravnavane kot večpomenske: *dimension* (razsežnost) – 6 pomenov, *sens* (smer) je razcepljen v dve enakozvočni gesli, *sens*[1] (občutek, smisel) in *sens*[2] (smer, dimenzija), slednji pa še v 3 pomene, *base* (osnova) ima 15 pomenov, *sommet* (vrh) 3 in *vertical* (vertikalen) 2. Teoretično in brez upoštevanja sobesedila bi takšno definicijo lahko razumeli na 1620 načinov! Slovaropisec ima seveda možnost, da z ustreznimi oznakami zagotovi enoznačnost leksemov. Takšna definicija bi lahko izgledala takole:[2]

> **HAUTEUR /.../** Dimension**I.1** dans le sens[2]**2** vertical**1**, de la base**I.1** au sommet**1**

2 Slovaropisci se za dodajanja posebnih alfanumeričnih oznak ne odločajo, saj bi po nepotrebnem otežile branje definicij. Edini slovar, ki je po našem vedenju apliciral takšno razločevalno sredstvo je Razlagalno-kombinatorni slovar (*Dictionnaire explicatif et combinatoire du français contemporain*) I . A. Mel'čuka, o katerem bomo spregovorili v drugem delu članka.

Posebno vprašanje nadalje je, ali je res naloga slovaropisca zajeti bistvo in naravo pojma, ne pa veliko skromneje, samo pojasniti pomene ali rabo posameznih enot besedišča (prim. Bolinger 1965).

Posebnost slovarske definicije je še očitnejša, če jo primerjamo z »definicijami« v filozofiji in različnih znanostih in vedah. V epistemologiji je definicija od Kanta naprej vedno razumljena v svoji konvencionalni razsežnosti. V znanostih in vedah definicija skupaj s poimenovanjem določi bistvene lastnosti določenega pojma, ki ga vpeljuje v diskurz (*z izrazom A imenujemo …*). V sodobni terminografiji definicija ne »predpiše« poimenovanja in bistvenih lastnosti nekemu pojmu, ampak neki ne dovolj jasno opredeljen pojem konvencionalno zameji in mu s tem omogoči enoznačnost. Slovarska definicija sodobnega slovaropisja pa nasprotno temelji na opazovanju in empirični in na korpusih temelječi analizi rabe posamezne leksikalne enote. V nemškem prostoru se tako že od sedemdesetih let raba termina *Definition* umika terminu *Explikation* v različnih besedotvornih različicah. V manjšem obsegu podobno velja tudi za angleško leksikografijo (*explanation* postopoma, a previdno nadomešča *definition*). V slovenskem slovaropisju je prevladal termin *razlaga*, ki ga med drugim uporabljajo snovalci SSKJ, čeprav po mojem vedenju nikjer ni bilo pojasnjeno, zakaj je *razlaga* primernejša od tradicionalno bolj uveljavljene *definicije* (prim. Gantar in Krek 2009).

V francosko govorečem prostoru konkurenčnega termina ni, se je pa že od sedemdesetih let dalje oblikovala celovitejša opredelitev slovarske definicije, zlasti seveda v odnosu do epistemološkega razumevanja definicije. Brata Dubois (Dubois in Dubois 1971: 84) izrecno poudarjata, da slovarske definicije ne moremo enačiti z znanstveno definicijo, niti ne s semantično analizo, saj omenjene oblike pripadajo različnim tipom diskurzov. Še bolj določna je J. Rey-Debove, ki poudarja, da je slovarska definicija tako po svoji funkciji kot skladenjsko-slogovnih značilnostih bližje »vsakodnevnemu«, »običajnemu« diskurzu. O slovarski definiciji govori kot o »naravni« definiciji.

> *Elle /la définition / fait partie d'un type de discours tout à fait ordinaire et particulièrement fréquent, qui est l'explication d'une pensée assurant le bon fonctionnement du dialogue*. (Rey-Debove 1971: 191)[3]

> *La définition est donc une activité naturelle et non métalinguistique dans son principe, qui répond à un besoin social primordial, celui de se faire comprendre. Le dictionnaire, en généralisant et en systématisant ce procédé, ne s'écarte pas beaucoup de la définition naturelle*. (Rey-Debove 1971: 192)[4]

3 /Definicija/ pripada povsem običajnemu in posebej pogostemu tipu diskurza, to je pojasnjevanje misli, ki omogoča nemoten potek dialoga.

4 Definicija je torej v svojem bistvu naravna in ne metajezikovna dejavnost, ki ustreza prvobitni družbeni potrebi, potrebi po medsebojnem razumevanju. Slovar se s splošno in sistematično definicijo tega postopka ne oddaljuje veliko od naravne definicije.

> /.../ *la définition est alors un énoncé comme les autres, une performance qui dépend de la comptétence linguistique ordinaire et non d'une science métalinguistique.* (Rey-Debove 1971: 197)[5]

Termin *naravne* definicije se je v (meta)leksikografiji ustalil in teoretsko razvil (glej tudi v nadaljevanju). R. Martin (1990) je izdelal tipologijo definicij, ki ločuje med konvencionalnimi in naravnimi definicijami. Konvencionalne definicije pripadajo znanostim ali vedam (*apriorne definicije*) in posameznim bolj ali manj institucionaliziranim terminologijam (*posteriorne definicije*), naravne definicije pa zajemajo različne oblike slovarskih definicij. Na pojem naravne definicije so se oprli tudi nekateri drugi slovaropisci in slovaroslovci. Wiegand (1989; 1992) npr. v »pomenskih parafrazah«, ki jih vsebujejo slovarji, vidi sistematizirane oblike vsakodnevnih naravnih dialogov o jezikovnih izrazih.

Ne glede na to, da sem osebno bolj naklonjen uporabi termina (*pomenska*) *razlaga,* bom v pričujočem članku zaradi navezave na specifičnost francoskega slovaropisja uporabljal termin *definicija.*

3 DEFINICIJA: OD IMPRESIONIZMA DO ZNANSTVENOSTI

Slovarska definicija je v zgodovini slovaropisja vezana predvsem na logično-filozofsko tradicijo, pri čemer je treba upoštevati, da se slovaropisci do sredine dvajsetega stoletja teoretskim vidikom pomenskega opisa niso posebej posvečali, ampak so se zanašali predvsem na svojo intuicijo in z leti slovarske prakse pridobljene »obrtniške« spretnosti.

3.1 Impresionistično obdobje

Definicija je do druge polovice 20. stoletja v slovaropisju predznanstveni pojem, o katerem se posebej teoretsko ni razpravljalo. Pri slovaropiscih so bili v ospredju pragmatični vzgibi in v primeru splošnih francoskih enojezičnih slovarjev zlasti cilj, da se z definicijami zajame in s tem predpiše t. i. »bon usage« (lepo, pravilno rabo), torej jezik vladajočih slojev, »legitimni« jezik (fr. *langue légitime*), kot bi rekel Bourdieu (1982: 13–58). Do 20. stoletja je predpisovalna vloga definicije pogosto prevladujoča. Definicije imajo tako nujno konvencionalno razsežnost, saj je njihova naloga normativno določati pomenski obseg leksemov. Vzemimo kot primer slovar Francoske akademije (prva izdaja 1694, danes se pripravlja deveta).

5 /…/ definicija je torej izrek kot vsi drugi, raba, ki je odvisna od običajnega jezikovnega znanja, in ne od metajezikovne znanosti.

Francoska akademija je bila ustanovljena predvsem z namenom izdati slovar, ki bi po zgledu slovarja toskanske Accademia della Crusca (prva izdaja 1612) prispeval k poenotenju in prečiščenju jezika. V Franciji je bil slovar pomemben politični projekt nastajajoče absolutistične države. Pomembno državotvorno, oziroma »državljanotvorno« vlogo je ohranil tudi v republiki.

Na razvoj slovaropisja v 17., 18. pa tudi še v 19. stoletju je pomembno vplivala racionalistično-razsvetljenska miselnost (šola Port-Royala, »enciklopedisti«, Voltaire, Bacon, Leibnitz) in sočasni razvoj znanosti in tehnike. Slovarji tistega časa (Nicot 1606; Richelet 1680; Furetière 1694; Féraud 1787) so hibridne oblike, ki združujejo prvine splošnih slovarjev, specializiranih terminoloških slovarjev in enciklopedij. Definicija je razumljena kot zdravilo, ki naj v naših mislih in govoru odpravi zmedo, ki jo tja zaseje nejasnost besed.

V drugi polovici 19. stoletja in z uvedbo obveznega šolanja se slovarji profilirajo kot pedagoška orodja (prvi »šolski« slovar izda P. Larousse leta 1856), ki naj vsakemu državljanu pomagajo, da se z znanjem standardnega jezika enakopravno uveljavi v družbi. Posledica razumljivo je, da sta tudi v samih definicijah še vedno v ospredju normativna in predpisovalna vloga. Hkrati se v tem času tudi v slovaropisju že čuti vpliv razvoja historične filologije. Takšen je primer slovarja E. Littréja (1863–1878), ki temelji na obsežnem korpusu skrbno izbranih citatov pisateljev zlasti 17. in 18. stoletja in ki vsebuje obsežno etimološko gradivo. Žal danes pogosto v njem vidimo predvsem poskus, kako s stališča historične normativnosti prispevati k ohranitvi neke idealizirane podobe francoskega jezika klasicističnega obdobja.

Definicija v tem »predmodernem« obdobju ni razumljena kot posebna besedilna ali diskurzivna zvrst, prav tako se slovaropisci sistematično ne posvečajo določitvi pomenskih prvin, ki naj jih vključijo v definicijo. Z današnjega stališča bi marsikateremu pomenskemu opisu težko pripisali status slovarske definicije.[6] Poglejmo dva primerov za besedo *chien*.

- Dictionnaire Universel avtorja A. Furetièra (1690)

> **CHIEN.** *s. m.* CHIENNE. *s. f.* Animal domestique qui abboye, qui sert à garder la maison, & à la chasse. Le *chien* est le symbole de la fidelité. Les *chiens* sont en telle abomination aux Maldives, que si un *chien* avoit touché quelqu'un du pays, il s'iroit incontinent baigner pour se purifier. Pyrard. Au contraire chez les Gaures ils sont en si grande veneration, que les Prêtres se servent des *chiens* pour purifier leurs penitents. Tavernier. Un *chien* fut

6 Neenotna strukturiranost slovarskih člankov in nepreglednost posameznih rubrik v starejših slovarjih je poseben problem, ki se posebej ostro zariše ob pripravi elektronskih različic (Wooldridge 1992).

établi pour Gouverneur de la Norvege par Gunnar Roy de Suede, aprés qu'il l'eut subjuguée, comme témoigne Saxo-Grammaticus. Ce mot vient du Grec *kyon, canis.* [7]

- 4. izdaja slovarja Francoske akademije (Dictionnaire de l'Académie française, 1762):

CHIEN, CHIENNE. s. Animal domestique qui aboie.[8]

S stališča sodobnega slovaropisja se definiciji približa edino akademijski slovar, pri prvem pa gre za mešanico slovarske definicije, o čemer posredno priča tudi navajanje morfoloških informacij, in enciklopedičnega opisa.

Posebno vprašanje je izbira pomenskih sestavin. A. Rey (1977: 114) je to »predmoderno« obdobje duhovito označil kot impresionistično, saj je temeljilo predvsem na intuiciji in introspekciji, ki nista bili v službi objektivnega opisa dejanske jezikovne rabe. Nesistematičnost se kaže tako v definicijah (izbira pomenskih sestavin, ubesedenje) kot v pomenski členitvi. Vzemimo primer besede *homme* (človek, moški) v slovarju E. Littréja. Beseda je členjena v 26 hierarhično nerazlikovanih pomenov. Osnovna težava je, da je temeljna pomenska členitev na *homme* – človek in *homme* – moški v množici pripisanih podpomenov zabrisana, saj se pomen *homme* – moški pojavi šele v 11. in 12. razdelku. Številne pomene, ki so navedeni v samostojnih razdelkih, večina uporabnikov s težavo intuitivno prepozna kot samostojne pomene.[9]

3.2 Slovaropisje kot uporabna veja jezikoslovja

Pomemben prelom je prinesla druga polovica 20. stoletja, ko je pričel razvoj sinhronega jezikoslovja odločilno vplivati tudi na slovaropisje. Uredniško vodstvo najpomembnejših francoskih založniških hiš so prevzeli že uveljavljeni jezikoslovci: J. Dubois in R. Lagane v založniški hiši Larousse, zakonca A. Rey in J. Rey-Debove pa v hiši Le Robert. V šestdesetih letih se je začel najbrž zadnji veliki institucionalni slovaropisni projekt v francoskem prostoru, Trésor de la langue française (TLF) (v 16 knjigah, 1971–1994).

V nadaljevanju sledi kratka predstavitev treh izstopajočih pogledov na slovarsko definicijo, ki sem jo členil v dva podrazdelka.

7 Slovarski opis zajema poleg klasične definicije (domača žival, ki laja, čuva hišo in je primerna za lov) tudi številne bolj ali manj naključno izbrane enciklopedične podatke o tem, kakšna je vloga in simbolna vrednost psa v različnih kulturah.

8 Smiselni prevod: pes – domača žival, ki laja.

9 Kljub obsežnim citatom težko uvidimo, v čem npr. je razlika med človekom kot najbolj razvito živaljo in človekom glede na lastnosti, ki ga dvigajo nad živali, ali med človekom kot človeškim bitjem na splošno, človekom kot pripadnikom človeške vrste in človekom kot posameznikom, ki pripada človeški vrsti?

3.2.1 Minimalna paradigma opisa leksikalnega pomena

Minimalna paradigma se je razvila z naslonitvijo na strukturalni model pomenske analize, zlasti na sestavinsko analizo (v ZDA J. Katz in J. Fodor, v Franciji B. Pottier, A. Greimas, F. Rastier, v slovenskem prostoru A. Vidovič Muha), ki v pomenu razločuje pomenske delce, seme. Sestavinska analiza temelji na Aristotelovem logično-spoznavnem modelu kategorizacije, ki ostro ločuje med bistvenimi in naključnimi lastnostmi, pri čemer so bistvene lastnosti tiste, ki določajo pripadnost nekega elementa človekovega izkustva določeni kategoriji. V sodobni semantiki takšen pristop imenujemo model nujnih in zadostnih pogojev (angl. *necessary and sufficient conditions*). Naloga slovarske definicije ni izčrpen opis leksikalnega pomena, ampak določitev pogojev denotiranja nekega leksema (Weinrich 1970: 73), kar obenem omogoči, da se neki leksem nedvoumno loči od ostalih.[10]

Prvi slovar, ki je poskušal dosledno aplicirati strukturalno metodologijo, zlasti Bloomfieldov distribucionalizem, je bil Laroussov Dictionnaire du français comtemporain (*DFC*) (1966) pod uredniškim vodstvom J. Duboisa. Pri pomenski členitvi so dosledno aplicirani sinhroni kriteriji: pomenski (pripadnost različnim leksikalnosistemskim paradigmam), oblikoslovno-skladenjski (besednovrstna pripadnost, vezljivost) in pretvorbeni (tvorjenje izpeljank, sestavljank, zloženk). Posledica je drobljenje makrostrukture. Leksemi, med katerimi kljub enaki etimologiji na sinhroni ravni ne moremo vzpostaviti pomenskih razmerij, so obravnavani v več ločenih člankih. Beseda *coeur* (srce) je npr. razdeljena v štiri ločene članke: *coeur* 1 – organ, *coeur* 2 – središče za čustva, *coeur* 3 – osrednji del stvari, *coeur* 4 – barva pri igralnih kartah. Druga posebnost slovarja na ravni makrostrukture je vključevanje besedotvorno povezanih besed (izpeljanke, sestavljanke, zloženke) v članek leksema, ki predstavlja osnovo besedotvornih operacij. Tako so npr. v članku *charger* (naložiti, natovoriti) gnezdeni še podčlanki *chargement* (natovarjanje, tovor), *décharger* (raztovoriti, razložiti), *surcharger* (preobtežiti).

V definicijah želi biti slovar ozko funkcionalističen, tako da se omejuje na navajanje samo tistih pomenskih sestavin, ki omogočajo minimalno razločevalnost. Poglejmo si primer definicije za *tournevis* (izvijač):

> *Outil pour serrer, desserer les vis*.

Definicija kot edino razločevalno prvino, ki izvijač ločuje od ostalih orodij, navede njegovo funkcijo (privijanje, odvijanje vijakov). Poglejmo še definicijo iz NPR, ki vsebuje še opis oblike in način delovanja.

10 O (ne)smiselnosti doslednega razumevanja definicije kot orodja za določanje jezikovnosistemske pomenske razločevalnosti se lahko prepričamo na naslednjem primeru, ki ga obravnava Weinrich (1970: 78). V definiciji leksema *carotte* (korenje) »*racine orange pointue commestible*« (užitna oranžna koničasta korenina) je odveč prvina *pointu* (koničast), saj v jezikovnem sistemu francoščine ni ubeseden pomen »*racine orange commestible non pointue*« (užitna oranžna nekoničasta korenina).

Outil pour tourner les vis, fait d'une tige d'acier emmanchée et présentant à son extrémité une forme s'adaptant dans l'empreinte de la tête de vis (en fente ou cruciforme).

3.2.2 Maksimalna paradigma opisa leksikalnega pomena

Že kmalu so se pojavile kritike minimalne paradigme leksikalnega pomena. J. Rey-Debove, tudi urednica pri založniški hiši Le Robert, je zelo kritična do prenosa strukturalne semantike v slovaropisje in dvomi o možnosti, pa tudi smiselnosti omejevanja definicije na minimalno število relevantnih pomenskih sestavin (Rey-Debove 1971: 213–218). V navajanju pomenskih sestavin, ki presegajo zahteve minimalne razločevalnosti in ki so s stališča strukturalnega pomenoslovja »odvečno natančne«, ne vidi pomanjkljivosti, ampak prej odliko slovarske definicije. Definicija, ki bi se omejila samo na minimalno število pomenskorazločevalnih prvin, bi od uporabnika zahtevala večji miselni napor in bi se lahko spremenila v uganko. Ker definicija pripada naravnemu jeziku, je določena mera redundantnosti nujna.

A. Rey (1977: 98–113), prav tako urednik pri založbi Le Robert, zagovarja stališče, da mora slovarska definicija zajeti psihosocialno resničnost govorcev nekega jezika, kot se ta odseva v besedah, in se ne sme omejevati samo na jezikovnosistemske razločevalne prvine. Potrebno je zajeti tudi bolj ali manj »naključne« informacije, ki se tičejo družbe, kulture ali zgodovine neke jezikovne skupnosti. Seveda pa je treba na drugi strani paziti, da definicije ne bi vsebovale informacij, ki nimajo širše družbene veljavnosti. Poglejmo si primer definicije za *antre* (duplina) iz NPR:

Caverne, grotte (spécialt. servant de repaire à une bête fauve).

Definicija je v prvem delu minimalna, saj se omeji na navedbo dveh sopomenk. Zanimivejši je dostavek v oklepaju, ki prinaša informacijo, kakšne konotacije oz. asociacije beseda zbuja francoskim govorcem in ki geselski leksem razlikujejo od obeh navedenih sopomenk (»ki služi divjim živalim kot brlog, skrivališče«).

Na kratko naj omenimo še eno posebnost NPR, to je vloga zgledov rabe, ki pogosto delujejo kot dopolnilo definicije. Vzemimo primer leksema *jardinet* (majhen vrt, vrtiček):

Petit jardin. *Les jardinets des pavillons de banlieue.*

Definicija je minimalna besedotvorno-pomenska definicija, ki ubesedi instrukcijski pomen pripone *-et*. Takšna definicija za razumevanje leksema v vsej njegovi

psihosocialni razsežnosti ni dovolj, saj francoski govorec vsakega vrta manjših dimenzij ne bi imenoval *jardinet*, zato je nujna navezava na manjše vrtove, ki obkrožajo individualna predmestna domovanja.

A. Rey in J. Debove-Rey (Rey-Debove 1989) sta maksimalno paradigmo pomenskega opisa v osemdesetih in devetdesetih letih še dodatno razvila in jo navezala na prototipsko teorijo (E. Rosch, G. Lakoff). Poleg razločevalnih lastnosti v smislu nujnih in zadostnih pogojev mora definicija podati lastnosti, ki sicer ne pripadajo celotni kategoriji, ampak samo posameznim prototipskim podvrstam, kategorijam ali entitetam. Takšne prvine so v NPR uvedene z veznikom, ki zameji njihovo splošno veljavnost (npr. *en général, généralement* ipd.). Poglejmo primera definicij za *oiseau* (ptič) in *boîte* (škatla, zaboj):[11]

> Animal appartenant à la classe des vertébrés tétrapodes à sang chaud, au corps recouvert de plumes, dont les membres antérieurs sont des ailes, les membres postérieurs des pattes, dont la tête est munie d'un bec corné dépourvu de dents, et qui est en général adapté au vol.
>
> Récipient de matière rigide (carton, bois, métal, plastique), facilement transportable, généralement muni d'un couvercle.

Prototipski pristop je delno apliciran tudi pri členitvi pomena.[12] V NPR je členitev hierarhična in se razvija na dveh do treh ravneh. Prva raven je pri kompleksnejših leksemih oblikoslovno-skladenjska (besednovrstna pripadnost, vezljivost), drugi dve sta pomenski in se členita na osnovne pomene in podpomene (rezultati pomenskih širitev, oženj, metaforičnih ali metonimičnih prenosov).

Zanimiv je primer že omenjenega TLF. Slovar je kljub enoviti zasnovi v več kot trideset letih svojega nastajanja[13] z menjavami v ekipi sprotno nekoliko spreminjal svojo prvotno uredniško politiko. Začetek projekta je vezan na strukturalno jezikoslovje. Prvi urednik, P. Imbs, v uvodu k prvi knjigi projekt opiše tudi kot poskus prenove tradicionalnih slovaropisnih metod pod vplivom strukturalne semantike in definicijo opredeli kot tradicionalno slovarsko obliko sestavinske analize, ki je v skladu z aristotelovskim razumevanjem definicije kot izreka, ki podaja najbližji rod in specifično razliko. Definicija vsaj v prvih zvezkih in na deklarativni ravni ostaja znotraj minimalne paradigme. V nadaljevanju, zlasti potem, ko uredništvo prevzame B. Quemada, se pomenski opis bolj nasloni na

11 Obe definiciji podata obče veljavno definicijo v smislu nujnih in zadostnih pogojev, za veznikom (*généralement, en général*) pa navedeta tudi prototipsko lastnost: za *ptiča*, da običajno leti, za *škatlo* pa, da ima običajno pokrov.

12 Veliko dosledneje je prototipski pristop apliciran v The New Oxford Dictionary of English (1998), kjer za razliko od NPR diahroni razvoj geselskih leksemov ne igra nobene vloge.

13 Priprave (nakup ustrezne infrastrukture, izdelava informacijskih orodij, oblikovanje korpusa, zasnova) so se začele v šestdesetih letih in končale s 16. knjigo 1994. Leta 1997 je bila izdana še dodatna knjiga z dopolnili in popravki. Projekt se je nadaljeval z informatizacijo (2004) in prenosom na splet.

teoretski model psihomehanike G. Guillauma, ki je v jezikovnih pojavih videl odraz psihičnih procesov. Členitev pomena in posamezne definicije naj odsevajo analogije med predvidljivimi pomenskimi spremembami. Pomenska členitev je kompleksna in obsega pet ali celo šest hierarhičnih ravni. Ključna je navezava na Guillaumov koncept »potencialnega označenca« (fr. *signifié de puissance*), katerega opis mora vsebovati prvine, ključne za razumevanje izpeljanih pomenov, ki so posebej definirani v pomenskih podrazdelkih. Poglejmo si osnovno definicija leksema *corbeau* (krokar):

> Grand oiseau (Passereaux) au plumage noir, au bec fort et légèrement recourbé, réputé charognard.

Definicija je zajela tri pomenske sestavine, na katere se navezujejo izpeljani pomeni (črna barva /*plumage noir*/: osebe črne polti, duhovnik; oblika kljuna /*bec fort et légèrement recourbé*/: posebno orodje, oblika nosu; mrhovinar /*charognard*/: grobar, brezvesten zaslužkar).

Pri nastajanju slovarja TLF je od začetka deloval tudi R. Martin, ki je v osemdesetih in v začetku devetdesetih metodologijo slovarskega opisa pomena navezal na teorijo stereotipa H. Putnama (Martin 1990, 1991). Pomembno je, da slovarska definicija pri uporabniku vzbudi ustrezno mentalno predstavo in da ima čim večjo obvestilno in pojasnjevalno vrednost. Po Martinu bi moral dober slovarski pomenski opis zajeti:

- univerzalne lastnosti, ki jih delijo vse z neko leksikalno enoto poimenovane entitete (za *ptiča* npr. kljun, perje, krila kot zgornje okončine);
- obče sprejete stereotipne lastnosti, ki jih poseduje večina poimenovanih entitet (za *ptiča* npr. da leti, poje …);
- simbolne stereotipne lastnosti, ki jih neka jezikovna skupnost pripisuje entiteti (za *ptiča* v francoščini npr. veselje, svobodnost, čudaštvo, skromne prehranjevalne navade …).

Žal R. Martin nikoli ni prevzel uredniškega vodenja slovarja TLF,[14] tako da so njegovi predlogi, kljub obsežnemu empiričnemu delu, ki ga je opravil kot univerzitetni profesor, ostali na ravni slovaroslovne teorije.

14 Slovar je vsaj kar zadeva pomensko analizo prepogosto ujetnik eklektičnega, a vseeno ozkega strukturalno-psihomehaničnega modela. Tako npr. najdemo tudi definicije povsem virtualnih in v korpusu nezabeleženih pomenov, ki pa so metodološko nujne, saj predstavljajo t. i. *signifié de puissance*. Beseda *bloc* (1) (blok, klada) ima kot osnovni pomen navedeno *»action de bloquer, de rendre immobile«* (*dejanje blokiranja, ustavitve*). Pomenski razdelek se sicer sklicuje na 8. izdajo slovarja Francoske akademije in je označen s kvalifikatorjem *rare* (redko), a beseda v korpusu Frantext, na katerem temelji slovar, tega pomena nima zabeleženega, prav tako ni pomen intuitivno dostopen govorcem, ki sem jih neformalno intervjuval (v tem pomenu se uporablja *blocage* - blokiranje).

4 DEFINICIJA: MED SLOVAROPISJEM IN SLOVARSTVOM

Razvoj informacijskih tehnologij, ki naglo in nezadržno prodirajo v slovaropisje, in bogata in raznolika slovarska produkcija v sedemdesetih in osemdesetih letih sta privedli do vpeljave koristnega metodološkega razlikovanja (Quemada 1987) med slovaropisjem (fr. *lexicographie*) in slovarstvom (fr. *dictionnairique*). Slovaropisje ustreza »predslovarski« fazi in zajema oblikovanje korpusov, njihovo obdelavo, analizo leksikalnih enot, pri čemer je končni produkt leksikalna podatkovna zbirka, ki ni neposredno namenjena trženju oz. izdaji konkretnega slovarja. Slovarstvo je priprava konkretnega slovarskega izdelka določenega obsega, z določenim proračunom, ki je običajno tržen za določeno ceno, ki je namenjen določenemu tipu uporabnikov ipd. V ospredju slovarstva so zlasti pragmatična (profil uporabnika, njegove potrebe, načini in možnosti uporabe) in komercialna vprašanja (raziskave trga, finančna struktura projekta, prodaja ipd.).

S stališča slovarstva slovar ni nevtralen in obče veljaven opis leksikalne ravnine jezikovnega sistema, ampak je pripomoček, ki govorcem pomaga pri razreševanju težav, na katere naletijo med razvezovanjem (dekodiranjem) ali uvezovanjem (enkodiranjem) jezikovnih sporočil, ali odpravljanju jezikovne negotovosti. V tej luči tudi slovarska definicija ni splošnoveljaven in izčrpen opis pomena leksikalnih enot, ampak odgovor, ki ga v namišljenem dialogu slovar posreduje na predvideno vprašanje uporabnika. Zaradi heterogenosti uporabnikov in njihovih potreb tudi ne more biti enega samega recepta za dobro definicijo. Prav tako moramo upoštevati pomensko heterogenost besed, saj npr. slovničnih besed ne moremo opisati na enak način kot polnopomenskih, pa tudi polnopomenske lahko opišemo na različne načine, ne da bi pogosto vedeli, katera oblika definicije je za koga in za kaj ustreznejša. Poglejmo si primera definicije za *appartement* (stanovanje) iz NPR in Laroussovega Dictionnaire du français d'aujourd'hui:

> Partie de maison composée de plusieurs pièces qui servent d'habitation.
>
> Local d'habitation, composé de plusieurs pièces contiguës, dans un immeuble qui comporte plusieurs de ces locaux.

Definicija iz *NPR* stanovanje definira po načelu del-celota (del hiše, ki …), medtem ko Laroussova v ospredje postavi funkcijo predmeta, ki ga geselska beseda označuje (bivalni prostor, ki …).

Kot smo že povedali, definicija Aristotelovega tipa ni edina vrsta slovarske definicije, niti ni najpogostejša. J. Rey-Debove (1971: 180–257) je v prvi pomembnejši slovaroslovni monografiji na podlagi analize obstoječih slovarjev izdelala

tipologijo slovarskih definicij, ki je z manjšimi dopolnitvami veljavna še danes (prim. Kosem 2006).

Najpogostejša definicija je po Rey-Debove besedotvorno-pomenska, ki izkorišča besedotvorno kompetenco govorcev, je ekonomična, a žal nujno generira potencialno krožnost, kar je za uporabnika zlasti tiskanih slovarjev pogosto moteče.

DÉFERRAGE /…/ Action de déferrer ; son résultat.[15] (NPR)

DRAGUEUR /…/ FAM Personne qui drague.[16] (NPR)

ÉLOGIEUX /…/ Qui renferme un éloge, des éloges.[17] (NPR)

Klasična definicija Aristotelovega tipa, imenovana tudi hiperonimna oz. inkluzivna, je značilna zlasti za besede, ki se umeščajo v kompleksnejše taksonomične celote:

IMPALA /…/ Petite antilope *(bovinés)* des savanes d'Afrique du Sud-Ouest.[18] (NPR)

ESPADRILLE /…/ Chaussure dont l'empeigne est de toile et la semelle de sparte tressé ou de corde.[19] (NPR)

Navedba hiperonima je pogosto nenatančna in zelo splošna (vrsta, oblika, tip …):

BEURRÉ /…/ Sorte de poire fondante. (NPR)

BOUQUETIÈRE /…/ Sorte de vase. (NPR)

Definicije se opirajo tudi na meronimna medleksemska oz. konceptualna razmerja (del-celota):

ALGOLOGIE /…/ Partie de la botanique qui étudie les algues. (NPR)

CUISSARD /…/ Partie de l'armure qui couvrait la cuisse. (NPR)

Pogoste so tudi definicije, ki obsegajo samo navedbo sopomenk. To velja zlasti za lekseme, ki so socialnozvrstno zaznamovani (*fam*. – pogovorna zvrst, *littér*. – literarna, visoka knjižna zvrst):

15 Smiselni prevod: odstranitev podkve – dejanje, da se podkev odstrani; rezultat tega dejanja.
16 Smiselni prevod: zapeljivec – oseba, ki zapeljuje.
17 Smiselni prevod: poln hvale – kar vsebuje hvalo, pohvale
18 Smiselni prevod: impala – manjša antilopa (iz vrste goved) jugozahodnih Afriških savan.
19 Smiselni prevod: espadrilja – obuvalo, katerega zgornji del je platnen, podplat pa pleten.

SE DÉMERDER /…/ FAM se débrouiller (NPR)

PIPERIE /…/ LITTÉR tromperie, leurre (NPR)

Definicija se lahko opre tudi na protipomenskost:

INCONTRÔLABLE /…/ Qui n'est pas contrôlable

SE TAIRE /…/ Rester sans parler, s'abstenir de parler

Edini tip definicije, ki se oddalji od »naravnega« jezika, je metajezikovna definicija, ki se eksplicitno nanaša na jezikovni znak in ne na njegov denotat, in ki se uporablja zlasti za slovnične besede in besede z oslabljenim leksikalnim pomenom:

CAR /…/ Conjonction de coordination qui introduit une raison expliquant ce qui précède, qui justifie ce qu'on a dit.[20] (NPR)

AÏE /…/ Interjection exprimant la douleur, et par ext. une surprise désagréable, un ennui.[21] (NPR)

DRÔLEMENT /…/ sert à donner une valeur intensive à des adjectifs.[22] (DFC)

Tipologija J. Rey-Debove ni mogla zajeti dveh tipov slovarskih definicij, ki se pojavita šele leto po izidu monografije (glej v nadaljevanju). Gre za stavčne definicije in pa definicijske zglede, ki so značilni za del šolskih slovarjev (fr. *dictionnaires d'apprentissage*).

Stavčna definicija se od klasične definicije razlikuje po tem, da je geselski leksem vključen v *definiens*. Poglejmo si primera iz Laroussovega slovarja Maxi débutants (Larousse MD). V prvem primeru gre za stavčno definicijo delno strokovne besede (*obèse*) in ki se kot pridevnik uporablja za ljudi s prekomerno telesno težo, v drugem za definicijo bivše španske valute (definicija še doda, da je pezeto zamenjal evro):

obèse /…/ Une personne **obèse** est une personne très grosse.

peseta /…/ La **peseta** était l'unité monétaire de l'Espagne. Elle a été remplacée par l'euro.

20 Smiselni prevod: kajti – prirednі veznik, ki uvaja razlog, ki pojasnjuje, kar je bilo rečeno, ki pojasnjuje, kar je bilo povedano.
21 Smiselni prevod: au – medmet, ki izraža bolečino, in širše, neprijetno presenečenje, nevšečnost.
22 Smiselni prevod: izjemno – služi za izražanje visoke vrednosti pridevnikov.

Stavčne definicije se pogosto razvijejo v definicijske zglede, kjer je geselski leksem vključen v manjše poljubno sobesedilo in praviloma pojasnjen s krajšo razlago:

> **pétanque** /…/ *Jean et Pierre font une partie de* ***pétanque***, de jeu de boules.[23] (Larousse MD)

> **persil** /…/ *M. Poulain met du* ***persil*** *dans la salade de tomates,* une plante aromatique.[24] (Larousse MD)

V nadaljevanju bomo v dveh ločenih razdelkih analizirali, kako lahko na slovarsko definicijo na eni strani vpliva profil uporabnika, na drugi pa specifična funkcija slovarja.

4.1 Definicija v šolskih slovarjih

Šolski slovarji (fr. *dictionnaire d'apprentissage*) so v francoskem prostoru slovarji, namenjeni šolarjem, katerih materni jezik je francoščina,[25] in sicer za različna starostna obdobja, od 3 let (vstop v t. i. *école maternelle*) do vstopa v gimnazijo, kjer dijaki postopoma pričnejo posegati po splošnih enojezičnih slovarjih. Trg je finančno donosen, tako da na njem vlada velika konkurenca. Zastopane so vse velike založniške hiše (Larousse, Hachette, Nathan, Le Robert, Bordas).

Šolski slovarji so dober primer odločilne vloge, ki jo pri pripravi slovarja igra slovarstvo. Tako je npr. izbor geslovnika v celoti podrejen kurikulumu za posamezne razrede, kar ima za posledico sorazmerno visok delež besed s področja slovnice, biologije, matematike, zgodovine in geografije. Slovarji praviloma vsebujejo obsežno slikovno gradivo, ki je prilagojeno posameznim starostnim stopnjam, prav tako pa tudi posebne večstranske preglednice zahtevanih znanj za posamezne šolske predmete. Elektronske različice[26] vsebujejo še zvočno gradivo (oglašanje živali, zvoki glasbenih instrumentov) in številne spletne povezave na dodatne aktivnosti in dodatne informacije. Slovarji se želijo uporabniku približati tudi tako, da poskušajo zakriti abstraktni značaj slovarskega besedila. V zgledih rabe in definicijah nastopajo imena ali pa so pisani v prvi osebi. V nekaterih slovarjih, npr. Petit Robert des enfants (1988), v primerih nastopajo stalni liki, ki zglede rabe povežejo v zgodbo, ki je predstavljena na koncu slovarja. V Dictionnaires des petits ours (1989) je slovar del zgodbe plišastih medvedkov, ki oživijo in pričnejo pisati slovar.

23 Smiselni prevod: petanka – Janez in Peter igrata petanko, balinata.
24 Smiselni prevod: peteršilj – G. Poulain paradižnikovi solati doda peteršilj, aromatično začimbo.
25 Ne smemo jih enačiti z *learner's dictionaries*, ki so namenjeni tujcem.
26 Prva elektronska izdaja šolskega slovarja je bil leta 1998 izdani CD-ROM slovarja Robert Junior.

Zgodovina šolskih slovarjev se v Franciji res začne že l. 1856 z Nouveau dictionnaire de la langue française P. Laroussa, ki je po obsegu in ceni namenjen šolarjem, v vseh ostalih elementih pa ostaja slovar, ki je dostopen predvsem odraslim. Žal so skoraj sto let šolski slovarji zgolj pomanjšave obsežnejših splošnih enojezičnih slovarjev z minimalnimi prilagoditvami.

Prvi pomemben napredek prinese l. 1949 slovar M. de Tora Larousse des débutants. Gre za prvi resni poskus, da se slovar tako na makro- kot mikrostrukturni ravni prilagodi učencem nižjih razredov osnovnih šol. Pri pomenskem opisu velja izpostaviti dve novosti. Ilustracije (1.500 ilustracij za geslovnik, ki obsega 18.000 besed) sistematično in smiselno dopolnjujejo definicije. Jezik definicij avtor prilagodi tako, da ne vključuje manj pogostih in zahtevnejših besed.[27] Pri tem se neposredno naveže na projekt »temeljne francoščine« (fr. *français fondamental*), ki ga pod pokroviteljstvom Unesca vodi G. Gougenheim. Gougenheim in ekipa na podlagi korpusa spontanih govorjenih besedil z upoštevanjem čiste frekvence izdelajo seznam 1.445 temeljnih besed na osnovni ravni in dodatni seznam 1.800 besed na zahtevnejši ravni.[28] De Toro je s skrbno izbiro definicijskega jezika, ki prioritetno vključuje »temeljno« besedišče, postavil normo, ki se jo trudijo upoštevati vsi šolski slovarji, čeprav konkretne rešitve pogosto bolj temeljijo na intuiciji slovaropisca, kot pa na konkretnih analizah (Pruvost 2001; 2003).

Drugi pomemben razvojni skok je prispeval l. 1972 izdani Dictionnaire du français vivant (Bordas) pod uredniškim vodstvom priznanega jezikoslovca M. Cohena. Slovar je prvi, ki vpelje stavčne definicije oz. definicjske zglede (glej zgoraj), sklicujoč se na Wittgensteinovo načelo, da besede nimajo pomenov, ampak poznajo samo rabe (Pruvost 2001). V naslednjih dveh desetletjih se je za takšno možnost pomenskega opisa odločila večina založniških hiš, zlasti za slovarje, nemenjene učencem nižjih razredov osnovne šole in predšolskim otrokom (tistim, ki obiskujejo t. i. *école maternelle*). Oblike pomenskega opisa so, kot smo nakazali že v prejšnjem razdelku, lahko zelo različne in sežejo od čisto kontekstualiziranih primerov brez dodatne pomenske razlage do klasičnih stavčnih definicij. Poglejmo si tri primere iz Dictionnaire en herbe (1989):

> **dent** /…/ *Ma **dent** est tombée, la petite souris va passer.* [29]

27 De Torov slovar je v tem pogledu predhodnik Longmanovega Dictionary of Contemporary English (1978), ki je v angleško govorečem svetu prvi vpeljal t. i. nadzor nad definicijskim jezikom za pedagoške slovarje, namenjene tujejezičnim govorcem. Za širši pogled prim. Neubauer (1987).

28 Slovar Dictionnaire fondamental G. Gougenheima je bil kot končni rezultat projekta izdan l. 1958. Slovar ostaja danes zanimiv predvsem s teoretskega stališča, njegova uporabna vrednost pa je zanemarljiva. Nekaj več kot 160 posnetih pogovorov ni moglo zagotoviti uravnoteženega korpusa, tako da je geslovnik slovarja tematsko in socialnozvrstno neuravnotežen. Še dodatna težava je bila preveč »premočrtna« uporaba slovarja v pedagoške namene, ki je v marsikaterem razredu pri pouku francoščine nehote vodila do siromašenja leksikalne kompetence učencev (Pruvost 2001: 87).

29 Smiselni prevod: zob – izpadel mi je zob, ponj bi prišla miška.

déménager /…/**Déménager**, c'est changer de maison ou d'appartement.[30]

déjeuner /…/ Le repas de midi s'appelle le **déjeuner**.[31]

V nekaterih slovarjih je poudarjen ludistično-poetični princip, s katerim želijo uredniki jezik in uporabo slovarja prikazati kot igro in zabavo. Poglejmo dva primera, vzeta iz *Le dictionnaire des enfants* (1991), ki pomen bolj nakažeta, kot res opišeta. V prvem je pomen pridevnika *plat* (nizek, raven, ploščat) opisan z nizom pogostih kolokacij (krožnik, mošnja, pete, pnevmatike) in se konča s pogovorno izpeljanko (*raplapla*), ki lahko pomeni sploščen, pa tudi brez energije, je pa v sami definiciji prisoten predvsem zaradi učinka rime. V drugem primeru je pomen pridevnika *magnétique* (magneten) opisan s priložnostno pesmico, ki govori o tem, kaj magnet lahko privlači in kaj ne (geselska beseda se v »stavčni« definiciji ne pojavi).

plat /…/ Assiette creuse, assiette **plate**, bourse pleine, bourse **plate**, talons hauts, talons **plats**, pneu à **plat** et raplapla.

magnétique /…/ Un aimant, c'est magnifique.
Ça n'attire pas les briques
mais ça attire les clous.
Un point, c'est tout!

V osemdesetih letih postajajo kritike stavčnih definicij oz. definicijskih zgledov vedno glasnejše in prihajajo tako s strani slovaropiscev kot učiteljev. Čeprav takšnim pomenskim opisom ni moč odrekati didaktične vrednosti, saj so bliže naravnim oblikam komunikacije in so zato manj abstraktne in laže razumljive, pa je težava takšnih definicij, da so pogosto premalo natančne in ne zajamejo ustreznega pomenskega obsega besede. Poleg tega je navezava na določeno sobesedilo ovira, da bi uporabnik pomen posplošil in tako besedo uporabil tudi v drugih sobesedilih.[32] V zadnjih dveh desetletjih se večina šolskih slovarjev vrača k tradicionalni definiciji, ki ji sledi zgled rabe.

4.2 Specifične oblike definicij v slovarjih kot pripomočkih za sistematično učenje besedišča

Prototipska vloga enojezičnih slovarjev je bila vedno razvezovalna, od tu tudi ključna vloga, ki jo v enojezičnih slovarjih igra definicija. Seveda so slovarji vedno

30 Smiselni prevod: preseliti se – preseliti se pomeni, da zamenjaš hišo ali stanovanje.

31 Smiselni prevod: kosilo – opoldanski obrok imenujemo kosilo.

32 Stavčne definicije v francoskem prostoru nikdar niso prodrle v slovarje za odrasle, kot je bilo to npr. v Collinsovem Cobuild English Dictionary for advanced Learners (1987). Kljub odobravanju pomembnega dela strokovne javnosti se praksa stavčnih definicij ni razširila. Zanimivo je primerjati kritike, ki letijo na račun Collinsovega slovarja (Rundell 2006), s kritikami, ki so letele na račun stavčnih definicij v francoskih šolskih slovarjih.

vključevali tudi informacije, ki so bolj kot za razvezovanje potrebne za uvezovanje (zlasti zgledi rabe, podatki o vezljivosti, družljivosti ipd.), vendar pa se je razmislek o enojezičnem slovarju kot pripomočku za uvezovanje začel razvijati šele ob koncu 20. stoletja (Béjoint 2007). Hkrati so se začeli pojavljati tudi slovarji, ki niso več namenjeni samo občasni rabi in za razreševanje konkretnih jezikovnih zagat, ampak sistematičnemu učenju besedišča in ki so namenjeni tako maternim kot tujejezičnim govorcem. Predstavili bomo dva takšna »slovarja«.

Prvi je Lexique actif du français I. Mel'čuka in A. Polguèra. Slovar se umešča v teoretski okvir Pomen-Besedilo (Steele 1990), ki ga je I. Mel'čuk s sodelavci razvil na Montréalski univerzi, zlasti pa se navezuje na osrednje orodje modela, to je razlagalno-kombinatorni slovar.[33] Funkcija razlagalno-kombinatornega slovarja je izčrpen opis pomenskih, oblikoslovnih, besedotvornih, skladenjskih in kolokacijskih značilnosti leksemov, ki mora uporabniku nuditi vse relevantne informacije, kako leksem razumeti ali uporabiti v različnih kontekstih. Lexique actif izkorišča samo del potenciala razlagalno-kombinatornega slovarja z namenom sistematičnega podajanja informacij, ki jih govorec potrebuje pri aktivni uporabi jezika, torej pri uvezovanju. Osrednja tipa informacij za geslovnik 386 leksemov sta t. i. pomenska izpeljava (mreža leksemov, ki so z geselskim v različnih vrstah medleksemskih, besedotvornih ali konceptualnih razmerij) in kolokacije. Slovarske definicije kot take ni. Nadomeščata jo kratka pomenska oznaka, ki ima predvsem pomensko razločevalno vlogo, in pa abstraktna shema, imenovana aktantska shema, ki geselski leksem vpelje v sobesedilo in kjer so v obliki spremenljivk in pomenskih oznak označeni tipični udeleženci (aktanti) procesa. Na spremenljivke se v besedilu članka navezujejo tudi elementi pomenske izpeljave in kolokacije. Poglejmo si nekaj primerov.

Članek, namenjen leksemu *avocat* (odvetnik), je v skladu s pomensko členitvijo razdeljen v dva dela, ki ju uvajata oznaki: *individu qui pratique un métier* (posameznik, ki opravlja poklic) in *individu qui a un certain comportement* (posameznik, ki se obnaša na nek način). Obe oznaki sta dopolnjeni še s kratkim izmišljenim zgledom rabe (*Il sera interrogé par la police en présence de son avocat* in *Il se fait l'avocat des sans-abris et des opprimés*). V primerih, kjer pomenske oznake ne morejo zajeti pomenskih odtenkov, so zgledi rabe nujni, saj imajo edini razločevalno vrednost. V članku *répugnance* (odpor) sta obe pomenski oznaki *sentiment négatif* (negativno čustvo) in šele zgleda rabe nam pomagata zajeti pomenski odtenek, ki je za prvi pomen bolj v smislu odpora, studa (*Comment*

33 Do sedaj so za francoščino izšli 4 zvezki (Dictionnaire explicatif et combinatoire du français /1984, 1988, 1992, 1999/), za ruščino eden (1984), za nekatere ostale jezike pa po nekaj člankov. Slovar je prvi vrsti teoretsko orodje, njegova neposredna uporabna vrednost je majhna. Je pa metodološko zanimiv in z aplikacijo ustreznih prilagoditev v fazi slovarstva uporaben tudi za širši krog uporabnikov. Na Mel'čukovo metodološko zasnovo sta se npr. oprla tudi S. Verlinde in J. Binon pri izdelavi Dictionnaire d'apprentissage du français langue étrangère ou seconde, ki je danes vključen v večjo leksikalno podatkovno zbirko (https://ilt.kuleuven.be/inlato/, dostop 1. 8. 2015).

cacher la répugnance que nous inspire ce type de nourriture), za drugi pa bolj v smislu odlašanja, pomisleka (*Ils ont manifesté de la répugnance à s'engager dans ce commerce douteux*).

Aktantska shema predstavi tipično situacijo, ki jo leksemi označujejo, in udeležence te situacije. Podane so tudi vezljivostne posebnosti (predlogi, možnost uporabe zaimkov ...). Za oba pomena *avocat* sta shemi:

> Individu X, qui est l'avocat de la personne Y [= de N, A_{poss}] pour Z.[34]
>
> La personne X est l'avocat de la personne Y [= de N, A_{poss}] auprès de la personne Z [auprès de N, devant N].[35]

V nadaljevanju so navedene informacije o najpomembnejših medleksemskih razmerjih (sopomenke, protipomenke, nad- in podpomenke), tipična imena za udeležence (npr. *nom pour Y* client, *nom pour Z* cause, procès, poursuite) in kolokacije, kjer nastopa geselska beseda. V razlagalno-kombinatornem slovarju so kolokacije opisane s formaliziranim mehanizmom leksikalnih funkcij (abstraktnih pomenskih invariant, ki se vzpostavljajo med kolokatorjem in osnovo kolokacije). V *Lexique actif* jih uvajajo krajše pomenske etikete, kot npr. *[Y]être le client d'un A.* ([Y] biti stranka odvetnika), *[Y] agir pour devenir le client Y d'un A.* ([Y] delovati, da bi postal stranka Y odvetnika), *Endroit où travaille l'A.* (kraj, kjer deluje odvetnik).

Dictionnaire du français usuel je nastal pod uredniškim vodstvom J. Picoche in se naslanja na njen lasten teoretsko-metodološki okvir, ki je izšel iz že omenjene psihomehanike G. Guillauma. Slovar v 442 člankih, posvečenih tematsko ključnim leksemom ali leksemskim parom obravnava 15000 najpogostejših francoskih leksemov, dobljenih na podlagi frekvenčne analize korpusa Frantext. Namen je sistematično nadgrajevanje znanja besedišča. Definicij ni, ampak jih nadomešča kratek izmišljeni zgled rabe in pomenska shema, ki podobno kot v Lexique actif vsebuje spremenljivke, ki označujejo udeležence procesa, ki ga leksem predpostavlja. V nadaljevanju sledi opis pomenskega in sobesedilnega obnašanja geselskega leksema, ki zariše mrežo preko različnih analoških razmerij povezanih leksemov, ki so tipografsko poudarjeni (velike tiskane črke). 15000 leksemov torej ni deležnih slovarskih definicij, ampak je potrebno njihov pomen in rabo izpeljati iz analoške mreže, ki se riše okrog tematsko ključnih leksemov. Poglejmo si del članka glagola *supposer* (domnevati). Članek je razdeljen v dva sklopa. Prvi se začne z zgledom rabe *Le général suppose que les renforts vont arriver*« (General domneva,

34 Smiselni prevod: posameznik X, ki je odvetnik osebe Y [= nekoga, svojilni pridevniški zaimek] za Z.

35 Smiselni prevod: oseba X je odvetnik osebe Y [= nekoga, svojilni pridevniški zaimek] pri osebi Z [pri /poljuben samostalnik/, pred /poljuben samostalnik/.

da bodo prišle okrepitve). Na zgled rabe se v nadaljevanju navezuje tudi pomenski opis. Za vsak podsklop sledi aktantska shema ali krajša pomenska oznaka. V članku *supposer* je prva shema *A1 suppose A2 qu'il ne sait pas* (A1 domneva A2, da ne ve). Sledi opis in zgledi rabe, ki vključujejo tudi stalne besedne zveze, ki so posebej označene (¶).

> Il suppose que A2 phrase à l'ind.: A1, ayant un PROBLÈME à RÉSOUDRE, en CONNAIT certaines DONNÉES, mais pas toutes. Il voudrait bien savoir si A2 a eu lieu dans le passé / a lieu actuellement / aura lieu dans l'avenir. Étant donné ce qu'il SAIT, il fait une SUPPOSITION concernant A2. Telles et telles CONDITIONS étant réalisées, A1 peut PENSER LOGIQUEMENT que A2 est plus PROBABLE que son contraire. Mais de toutes façons, A2 est DOUTEUX.
> *Supposez que vous ayez un accident, ¶ dans cette supposition, syn. en langage un peu vulgaire, Une supposition, que vous ayez un accident, syn. SI vous aviez un accident, qu'arriverait-il à vos enfants? Il faut prendre une assurance! - C'est une simple supposition, une supposition GRATUITE : je n'affirme pas sa probabilité, mais tout peut arriver. - ¶ À supposer même que j'aie un accident, ils sont couverts, je suis bien assuré.*
> Syn. A1 FORMULE une HYPOTHÈSE. ¶ *Dans l'hypothèse où vous auriez un accident, que deviendraient votre famille? - Cet accident est purement HYPOTHÉTIQUE.*
> *Le général suppose les renforts arrivés:* il imagine sa supposition réalisée et envisage ce qu'il ferait dans ce cas.
> Emploi affaibli de supposer. Je suppose que tu as bien réfléchi, que tu n'as rien oublié avant de partir, etc. ...: je l'IMAGINE, je le CROIS.

Oba omenjena slovarja, Lexique actif in Dictionnaire ususel, v uvodu predlagata različne tipe pedagoških aktivnosti, ki jih učitelji s slovarjema lahko izvajajo v razredu.

5 ZAKLJUČEK

Sodobni razvoj francoskega slovaropisja, slovarstva in teoretskih slovaroslovnih raziskav je potrdil prepričanje, da je slovarska definicija posebna besedilna in diskurzivna zvrst, ki je v službi slovarja in njegovih uporabnikov, in ne občeveljavni znanstveni opis pomena neke besede. Slovarske definicije so bile v preteklosti pogosto deležne številnih kritik, ki so nemalokrat izhajale iz napačnega razumevanja njihove funkcije. Kot smo videli zlasti v 4. razdelku, so slovarske definicije v posameznih vrstah slovarjev lahko nadomeščene z drugačnimi vrstami pomenskega opisa, ne da bi bil pri tem uporabnik nujno prikrajšan.

Z nastopom korpusnega jezikoslovja in elektronskih slovarjev in slovarskih pripomočkov se je zdelo, da definicije ne bodo več potrebne (Béjoint 2007), saj jih bodo nadomestili primeri avtentične rabe. S slovaropisnega stališča so bila taka predvidevanja razumljiva, s slovarstvenega pa so nekoliko manj, saj je težko na pleča uporabnika prelagati določanje pomenskega obsega in rabe posamezne besede.

Vsaj kar zadeva slovarske definicije in pomenski opis, elektronska slovarska orodja še niso ponudila bistvenih novosti. Eden od razlogov je, da so se informacijske tehnologije doslej bolj posvečale slovaropisju in znatno manj slovarstvu. Od elektronskih slovarskih orodij v prihodnosti pričakujemo, da bodo bolj polivalentna in da bodo vključevala možnost, da se tudi pomenski opisi po želji prilagajajo uporabnikom in njihovim potrebam (Galisson 2001; Béjoint 2007).

Slovarski zgledi

Iztok Kosem

Abstract
In this paper, the role of examples in dictionary entries is presented, and an overview provided of relevant studies into the use and usefulness of examples. We put forward the different ways of presenting examples in general monolingual dictionaries, list the characteristics of a good dictionary example, and discuss the different methods of finding good examples. The focus then turns to the role and characteristics of examples in the proposal for a dictionary of contemporary Slovenian, the methods for their extraction, and the procedures to be followed for saving examples to the dictionary database and archiving them, before concluding with the different visualisation options for the (online) dictionary.

Keywords: dictionary examples, good examples, automatic extraction, visualisation, dictionary database

Ključne besede: slovarski zgledi, dobri zgledi, avtomatsko luščenje, vizualizacija, slovarska baza

1 UVOD

Zgledi so eden pomembnejših elementov slovarskega gesla, saj z njimi prikažemo rabo besed, njihovih kolokacij, stalnih zvez, frazeologije ipd. v kontekstu, torej kot se pojavljajo v dejanski jezikovni rabi. Upoštevajoč dejstvo, da večina jezikovnega opisa v slovarju dekontekstualiziranega, so zgledi kot kontekstualizirani primeri rabe besed za slovarske uporabnike ključnega pomena.

V prispevku najprej opredelimo vlogo zgledov v slovarju in podamo nekaj ugotovitev raziskav o pomembnosti slovarskih zgledov. Sledi pregled tujih in slovenskih slovarskih praks v enojezičnih slovarjih pri podajanju zgledov. Opredelimo tudi lastnosti dobrih slovarskih zgledov in predstavimo različne metode njihovega iskanja. Nato se osredotočimo na vlogo in lastnosti zgledov v predlaganem slovarju sodobnega slovenskega jezika (SSSJ), njihovemu pridobivanju iz korpusnega gradiva in beleženju v slovarski bazi, nekaj besed pa namenimo tudi vizualizaciji zgledov v slovarju. Zaključek je namenjen strnitvi razmišljanj in razmislekom o prihodnosti vloge ter beleženja slovarskih zgledov v slovarjih na splošno.

2 VLOGA ZGLEDA V SLOVARJU

Zgledi v slovarju opravljajo dve vlogi: receptivno in produktivno. Glavna receptivna vloga zgledov, ki je sicer tudi njihova temeljna vloga v slovarjih, je dopolnjevati slovarske razlage, zato morajo zgledi vsebovati predvsem s pomenom povezane informacije. Kot poudarjata Atkins in Rundell (2008: 454), uporabnik včasih brez zgledov težko razume razlago. Zgledi so lahko zelo koristni tudi pri navigaciji po (daljših) geslih, saj uporabniki lahko »prepoznajo pomen, ki ga iščejo, tako da poiščejo zglede z ubeseditvijo, podobno tisti, ki so jo prebrali ali jo želijo producirati« (Fox 1987: 137).

Produktivna vloga zgledov se nanaša predvsem na ponazarjanje skladenjskih vzorcev, vezljivosti, kolokacij in podobnih značilnosti iztočnice (Húmble 2001), ki naj bi uporabnikom pomagali pri pisanju ali (redkeje) govorjenju. Produktivno naravnane zglede najdemo predvsem v slovarjih za tiste, ki se jezika učijo, npr. slovarjih za tuje govorce ali slovarjih za mlajše materne govorce (šolskih slovarjih ipd.).

Raziskave, ki se ukvarjajo s slovarskimi zgledi, se osredotočajo predvsem na njihovo korist pri produkciji za nematerne govorce. Pri tem največkrat uporabijo metodo, pri kateri morajo udeleženci raziskave oblikovati stavke z (neznanimi) besedami, pomagajo pa si lahko s slovarskimi informacijami. Udeleženci so razdeljeni v več skupin: nekateri imajo na voljo samo definicije, drugi definicije in zglede,

v nekaterih raziskavah (npr. Frankenberg-Garcia 2012, 2014) je uporabljena še skupina udeležencev, ki ima na voljo samo zglede. Ugotovitve večine raziskav (Summers 1988; Laufer 1993; Nesi 1996; Al-Ajmi 2008) niso preveč spodbudne, saj kažejo, da zgledi ne prinesejo veliko dodane vrednosti uporabnikom pri produkciji. Vendar pa Frankenberg-Garcia (2012) poudarja, da imajo omenjene raziskave dve ključni metodološki pomanjkljivosti: kot prvo, čeprav merijo produktivno vrednost zgledov, so naloge zastavljene tako, da morajo udeleženci študije najprej razvozlati pomen neznane besede in jo nato takoj uporabiti v stavku, torej sta združeni receptivna in produktivna raba, kar je pri uporabi slovarjev redko. In drugič, uporaba neznanih besed pri preverjanju produkcije ne odraža realne rabe slovarjev in jezikovne produkcije nasploh, saj »ljudje pri pisanju redko uporabljajo njim povsem nove besede« (Laufer 1993: 138).

Frankenberg-Garcia (2012, 2014) je izboljšala metodologijo prejšnjih raziskav tako, da je jasno ločila preverjanje receptivne in produktivne vloge zgledov ter obenem ločila zglede na receptivno in produktivno naravnane. Udeležence raziskave je razdelila v štiri skupine: kontrolno skupino (brez slovarja), skupino, ki je dobila samo definicije, skupino, ki je dobila en korpusni zgled, in skupino, ki je dobila več korpusnih zgledov. Ugotovitve so pokazale, da je pri razumevanju pomena besede več korpusnih zgledov skoraj enako učinkovitih kot razlaga, pri produkciji pa ima več korpusnih zgledov precej večjo korist kot samo eden, zgledi pa so nasploh precej bolj koristni od razlag.

Raziskave, ki ponujajo informacije o tem, kako pogosto uporabniki pogledajo zglede, so redke. Béjoint (1981) je v svoji študiji s študenti ugotovil, da pri uporabi slovarja zglede pogledajo precej pogosto. O podobnih rezultatih poroča tudi raziskava med 620 študenti (449 maternimi in 171 nematernimi govorci angleščine) na univerzi Aston (Kosem 2010); zgledi so bili četrti najpogosteje uporabljeni del slovarskega gesla (za definicijami, izgovorjavo in sinonimi), pri nematernih govorcih pa celo drugi najpogosteje uporabljeni del gesla (za definicijami).

3 ZGLEDI V SPLOŠNIH ENOJEZIČNIH SLOVARJIH

Analiza splošnih enojezičnih slovarjev[1] kaže, da obstajajo glede na obravnavo in prikaz zgledov tri skupine slovarskih virov. V prvo skupino sodijo slovarji, ki zglede podajajo v obliki iztržkov (iz gradiva prevzetih in primerno abstrahiranih zvez), občasno poleg iztržkov ponudijo tudi celostavčne zglede (španski slovar)

1 Analiza je vključevala samo tuje slovarje, ki obstajajo v spletni obliki. Analizirani spletni slovarji imajo lahko tiskane različice oz. so izpeljani iz tiskanih različic, najdemo pa tudi slovarje, ki obstajajo samo v spletni obliki (npr. poljski, nizozemski). Slovenski slovarji so obravnavani ločeno po opredelitvi vseh treh skupin.

in/ali njihov vir (npr. estonski slovar), pri čemer nekateri slovarji zglede ponudijo samo pri določenih (pod)pomenih ali frazah. Gre predvsem za (nekorpusne) slovarje, ki so bili prvotno izdelani za tisk in potem ponujeni tudi na spletu, ali pa za nedavno objavljene slovarje, ki so bili konceptualizirani na osnovi leksikografskih praks prejšnjega stoletja. Med omenjene slovarje spadajo Slovar češkega knjižnega jezika[2] (Slovník spisovného jazyka českého, 1989), slovar španskega jezika Kraljeve španske akademije[3] (Diccionario de la lengua Española, 2014), Razlagalni slovar estonskega jezika[4] (Eesti keele seletav sõnaraamat, 2007) in Hrvaški enciklopedični slovar[5] (Hrvatski enciklopedijski rječnik, 2003).

Drugo skupino predstavljajo slovarji, v katerih prevladujejo celostavčni (korpusni) zgledi, iztržki so uporabljeni precej redkeje ali pa sploh ne. Primeri takšnih slovarjev so slovarji Oxford (Oxford Dictionaries[6]), Macmillan (Macmillan English Dictionary[7]) in Merriam-Webster (The Merriam-Webster Online Dictionary[8]) za angleščino, Slovar sodobnega danskega jezika (Den Danske Ordbog[9]), Veliki slovar poljskega jezika[10] (Wielki Słownik języka Polskiego) in Splošni nizozemski slovar[11] (Algemeen Nederlands Woordenboek). V omenjenih slovarjih zasledimo precej različne prakse navajanja zgledov: v Macmillanu in Slovarju sodobnega danskega jezika so celostavčni zgledi del prikazanega gesla, podobno velja za Merriam-Webster, kjer pa so celostavčni zgledi združeni pod ločeno rubriko proti koncu gesla, medtem ko so pri pomenih, podpomenih ipd. podani iztržki. Tudi slovar Oxford prvotno ponuja iztržke, celostavčni zgledi so pod vsakim pomenom oz. podpomenom na voljo s klikom gumba *More example sentences*. Poljski in nizozemski slovar zgledov sploh ne ponujata na prvi ravni prikaza rezultatov, ampak si jih uporabnik lahko ogleda s klikom na povezavo (nizozemski slovar) oz. zavihka (poljski slovar). Omenjena slovarja ter tudi Slovar sodobnega danskega jezika za vsak zgled navajajo tudi vir.

V tretji skupini najdemo portale, kot je nemški DWDS[12] (Das Digitale Wörterbuch der deutschen Sprache), ki uporabnikom na enem mestu ponuja zadetke iz slovarjev in korpusov ter druge relevantne informacije o iskani besedi.[13] Za to skupino virov je z vidika zgledov zlasti pomembna povezava med slovarji in korpusi, sploh z vidika ugotovitev Frankenberg-Garcie o koristih ponujanja

2 http://ssjc.ujc.cas.cz, (spletna verzija na voljo od 2011).
3 http://lema.rae.es/drae (dostop 8. 8. 2015).
4 http://en.eki.ee/dict/ekss (dostop 8. 8. 2015).
5 Dostopen prek Hrvaškega jezikovnega portala, http://hjp.novi-liber.hr (dostop 8. 8. 2015).
6 http://www.oxforddictionaries.com/ (dostop 8. 8. 2015).
7 http://www.macmillandictionary.com/ (dostop 8. 8. 2015).
8 http://www.merriam-webster.com (dostop 8. 8. 2015).
9 http://ordnet.dk/ddo (dostop 8. 8. 2015).
10 http://wsjp.pl (dostop 8. 8. 2015).
11 http://anw.inl.nl (dostop 8. 8. 2015).
12 http://www.dwds.de/ (dostop 8. 8. 2015).
13 Tudi nekateri ostali slovarji, kot je na primer danski slovar, na spletni strani ponujajo dostop do korpusa, vendar pa ne ponujajo enotnega iskanja po vseh virih in hkratnega prikaza zadetkov.

večjega števila zgledov uporabnikom. Slaba plat takšnih portalov je, da je uporabniku predočeno veliko različnih informacij,[14] kar otežuje njihovo interpretacijo in rabo.

Če se ozremo po slovenskih slovarjih, Slovar slovenskega knjižnega jezika (SSKJ) in njegov naslednik SSKJ2 sodita v prvo od zgoraj omenjenih skupin slovarjev, saj ponujata zglede v obliki iztržkov (temelječih na virih in tudi izmišljenih[15]). To za SSKJ ni presenetljiv podatek, saj je nastajal v času predkorpusne leksikografije. Dejansko je količina zgledov v geslih SSKJ precej obsežna in odstopa od marsikaterega podobnega tujega slovarja, tudi takšnih, ki so bili izdelani nedavno, kot je slovar španskega jezika Španske kraljeve akademije. Zgledi so tudi eden od slovarskih elementov v SSKJ, ki je pri pripravi SSKJ2 doživel večje spremembe, saj so jih avtorji popravljali in nadomeščali (zaradi sprememb družbenega sistema ipd.) ali pa so dodajali povsem nove zglede. A kot ugotavlja Krek (2014: 146), spremembe obstoječih zgledov pogosto niso ustrezne oz. potrebne ali pa novo dodani zgledi ne prinašajo neke nove dodane informacije za razumevanje pomena. Nadomeščanje oz. popravljanje obstoječih zgledov v SSKJ2, predvsem z razlogom sprememb družbenega sistema, se sploh zdi redundantno glede na to, da avtorji slovar predstavljajo kot vir, ki odseva 150 let slovenskega jezika.[16] To potrjuje tudi Krekov zaključek, da je bil v ponazarjalnem gradivu »izbrisan precejšen del resničnosti dobe pred letom 1991« (ibid: 147).

Še pred SSKJ2 je izšel Slovar novejšega besedja slovenskega jezika (2012; SNB), katerega avtorji so deloma upoštevali sodobne leksikografske trende in v gesla poleg iztržkov vključili (celostavčne) korpusne zglede. Kot je navedeno v Uvodu (SNB: 9), je bil glavni besedilni vir pri pripravi slovarja 318-milijonski korpus Nova beseda:[17]

> Na podlagi avtentične besedilne rabe, izpričane v tristomilijonskem besedilnem korpusu Nova beseda, smo v 5384 slovarskih sestavkih interpretirali 6512 pomenov in podpomenov aktualnih besed in besednih zvez, ki imajo svoj izvor v raznolikih področjih družbene dejavnosti.

Analiza zgledov pokaže, da so avtorji slovarja v odsotnosti (dobrih) zgledov črpali zglede tudi iz drugih korpusov, zlasti iz 1,2-milijardnega korpusa Gigafida. To samo na sebi ni sporno, postavlja pa pod vprašaj zgoraj citirano metodologijo izdelave geslovnika, sploh pri geslih, kot je npr. *bandži skok*:

14 DWDS sicer omogoča omejevanje zadetkov samo na določene vire.

15 Kot piše v Uvodu v SSKJ (1991: XXII), »/k/adar je bilo slovarsko gradivo pomanjkljivo, so naredili uredniki iztržke po drugih virih ali po spominu«.

16 Marko Snoj 2. novembra 2013 za STA: http://www.rtvslo.si/kultura/knjige/akademska-vojna-okrog-novega-slovarja/321592 (dostop 8. 8. 2015).

17 http://bos.zrc-sazu.si/s_beseda3.html (dostop 8. 8. 2015).

bándži skòk -- skôka in skóka m (ô, ȍ ó; ǫ)
skok v globino, pri katerem je skakalec pripet z dolgo elastično vrvjo; skok z elastiko: Obnaša se kot frkolin, ki se pred tovarišijo postavi z bandži skokom, ko se privezan na elastično vrv vrže z mostu v globel **E ↑bungee (jumping) in (↑)skòk**

Navedeni zgled je (rahlo) modificiran stavek iz korpusa Gigafida,[18] problematično pa je, da korpus Nova beseda ne vsebuje niti enega zadetka za *bandži skok* (v Gigafidi jih je 5). V tem primeru torej zgled ponazarja iztočnico, za katero sploh ne vemo, kako je prišla v slovar. Poleg tega so ravno zaradi osredotočenosti slovarja na novejše besedje, ki se praviloma v korpusih pojavlja redko, zgledi velikokrat omejeni zgolj na ponazarjalno vlogo, saj ne prinašajo dodane vrednosti k razumevanju pomena.

Bolj sistematičen (in popolni) korpusni pristop je bil uporabljen pri izdelavi Leksikalne baze za slovenščino (Gantar et al. 2012; LBS), ki vsebuje 2.500 gesel, v katerih najdemo 152.996 zgledov oz. povprečno več kot 61 zgledov na geslo. Vsi zgledi so celostavčni in izhajajo iz korpusa Gigafida (Logar Berginc et al. 2012). Zgledi v LBS niso modificirani, saj se izbira zgledov z vidika oblikovanja slovarske baze in izdelave slovarja razlikuje; v slovarski bazi namreč težimo k izbiri večje količine korpusnih zgledov, ki so tipično celi stavki, vzeti iz korpusa, in predstavljajo kandidate za dobre slovarske zglede (Gantar 2015). LBS je za slovenski, pa tudi širši prostor pomemben vir zaradi uporabljene metodologije. Med izdelavo LBS so bile namreč razvite in preizkušene številne metode, ki kombinirajo leksikografsko delo z avtomatskim luščenjem podatkov (tudi zgledov) ter predstavljajo temelje izdelave slovarja sodobnega slovenskega jezika in z njim povezane slovarske baze (gl. razdelek 5).

4 KAKŠEN JE DOBER SLOVARSKI ZGLED?

Med največkrat omenjenimi lastnostmi dobrih zgledov so naravnost oz. pristnost, tipičnost, informativnost in razumljivost. O pristnosti zgleda govorimo, ko deluje naravno, tj. ko je tak, kakršnega bi v jezikovni rabi dejansko srečali. Ravno zato se pristnost velikokrat povezuje tudi z avtentičnostjo, ki naj bi jo zagotavljalo izbiranje zgledov iz zbranih avtentičnih besedil oziroma korpusov, kar je v sodobni leksikografiji tako rekoč standard. Ob tem velja omeniti, da so že slovarji, ki so nastajali v času predkorpusne leksikografije, vsebovali zglede iz avtentičnih besedil (npr. the Oxford English Dictionary) ali vsaj na besedilih vsebovane dele zgledov oz. iztržke (npr. SSKJ). Vendar pa so takrat leksikografi zglede velikokrat oblikovali kar sami oz. so si jih izmislili na podlagi lastne intuicije; omenjeno prakso so korpusne študije (npr. Sinclair 1991; Hunston in

18 http://www.gigafida.net/ (dostop 8. 8. 2015).

Laviosa 2001) postavile pod velik vprašaj, zlasti za namene vključevanja zgledov v splošne enojezične slovarje.[19]

Sorodno načelu pristnosti je načelo tipičnosti, ki veleva, da morajo slovarski zgledi pokazati tipično rabo iztočnice z vidika konteksta, skladnje, frazeologije in kolokacij. Sodobna orodja za analizo besedil leksikografom pri iskanju tipičnih zgledov precej pomagajo, saj z njihovo pomočjo lahko poiščejo slovnične strukture, kolokacije in koligacijske lastnosti iztočnice (npr. v kateri obliki ali sklonu se beseda z določenim kolokatorjem najpogosteje pojavlja).

Informativen zgled daje slovarskemu geslu dodano vrednost, največkrat v povezavi z definicijo, ki naj bi jo uporabnik zaradi zgleda lažje razumel. Zgled hkrati potrjuje informacije, podane v definiciji, in kontekstualizira rabo iztočnice v določenem pomenu ali podpomenu. Informativnost je deloma povezana tudi s količino zgledov v posameznem geslu; elektronski mediji resda ponujajo možnost vključitve zelo velikega števila zgledov, a dejstvo, da se pri vsakem dodatnem zgledu pojavlja vprašanje, kaj res novega prispeva h geslu, kliče po preudarnosti njihove uporabe, če želimo ohraniti informativnost in uporabniško prijaznost slovarja. Po drugi strani pa ne gre pozabiti ugotovitev Frankenberg-Garcie (2012; 2014) o tem, da je več korpusnih zgledov včasih celo bolj koristnih od same definicije.

Razumljivost zgleda dosežemo z izogibanjem kompleksnim strukturam, redki in zahtevni leksiki ter pretirani dolžini. Na ta način se bodo uporabniki lahko osredotočili na iztočnico in od njih ne bomo zahtevali pretiranega miselnega napora za procesiranje informacij v sobesedilu. Seveda se določenim elementom težko izognemo za vsako ceno; npr. redkejše in »zahtevnejše« besede se rade pojavljajo skupaj z drugimi zahtevnimi besedami ali besednimi zvezami, ki jih moramo vključiti v zgled, če se želimo držati načela naravnosti in tipičnosti. Zgledi ne smejo biti predolgi, prekratki pa tudi ne, sploh če je slovar namenjen tudi za produktivno rabo, saj uporabniku v slednjem primeru ponudijo premalo informacij za uspešno produkcijo, pogosto pa tudi recepcijo.

Vse večji pomen ima tudi oblika zgleda, saj se je v sodobni leksikografski praksi uveljavila uporaba celostavčnih zgledov, tudi v splošnih slovarjih za materne govorce, ki so sicer še pred nekaj desetletji uporabljali skoraj izključno iztržke ali kratke zglede. Temu je nedvomno botrovala prevlada elektronskih medijev, zlasti spletnega, kjer ni potrebno pretirano varčevanje s prostorom, hkrati pa tudi spoznanja, da iztržki in podobni kratki zgledi, iztrgani iz stavkov oz. povedi, delujejo abstraktno in nenaravno (gl. Williams 1996).

19 Nekoliko v manjši meri to velja za slovarje za tuje govorce, saj kot navajata Atkins in Rundell (2008: 456), je veliko predkorpusnih angleških slovarjev za tuje govorce vsebovalo veliko dobrih slovarskih zgledov, ki so delovali avtentično, a niso bili vzeti iz avtentičnih besedil.

Posebna in nadvse pomembna tema pri izbiri zgledov je ideološki vidik, saj se predvsem v zgledih izkazuje slovarska ideološkost, torej resničnost, kot jo vidijo leksikografi. Z zgledi namreč leksikografi velikokrat povedo tisto, česar ne morejo v razlagi zaradi omejitev v definicijskem jeziku in tudi veliko večje eksplicitnosti glede ideološkosti (prim. Meschonnic 1991; Béjoint 2000; Epple 2000; Schutz 2002; pri nas Gorjanc 2004; 2005; 2012). Tako so zgledi dejansko mikrostrukturni element, v katerem se lahko najbolj odsevajo družbene vrednote, preko tega pa tudi izpostavljajo vrednostni sistem slovarske ekipe (Gorjanc 2014). Gorjanc (2014) na primeru izrazja v SSKJ, povezanega s homoseksualnostjo, ponazori, kako se lahko družbeni predsodki v slovarju zabeležijo kot sprejemljivi oz. del norme. Težave z ideološkimi spremembami pri zgledih rabe najdemo tudi v SSKJ2 (Krek 2014: 145–147). Leksikografi se torej morajo zavedati svoje nenevtralne vloge in pri izbiri zgledov (in seveda pri pisanju slovarja nasploh) ravnati družbeno občutljivo in odgovorno (Béjoint 2000: 124).

Najti zgled, ki bi ustrezal vsem naštetim kriterijem, še zdaleč ni enostavno. Čeprav imajo danes leksikografi na voljo zelo velike korpuse in posledično veliko potencialnih zgledov za posamezno iztočnico, se velikokrat zgodi, da najdejo stavke, ki izpolnjujejo dva kriterija, tudi tri, zelo redko pa stavke, ki izpolnjujejo vse kriterije dobrega slovarskega zgleda. Dejansko bi lahko kandidate za zglede razvrstili v stopnje od slabih, bolj slabih kot dobrih, dokaj oz. potencialno dobrih in dobrih, pri čemer naj bi bili dobri tisti, ki jih lahko iz korpusa prenesemo neposredno v slovar. A takšnih neposredno uporabnih dobrih zgledov je malo, precej več je potencialno dobrih, torej takšnih, ki bi ob malenkostnih popravkih postali dobri zgledi. Ampak če se odločimo v slovarju uporabljati modificirane zglede, kaj to pomeni za načelo avtentičnosti? Je naš slovar še korpusni? Kot pravita Atkins in Rundell (2008: 458), je pogosto omenjena izbira med izmišljenimi in avtentičnimi zgledi zavajajoča, saj ne odraža realne leksikografske prakse. Tudi slovarji, kot je COBUILD, izdelani s popolnim korpusnim pristopom, vključujejo modificirane zglede, čeprav velja poudariti, da so se avtorji slovarja COBUILD skušali tej praksi, če je bilo le mogoče, izogniti (Fox 1987).

Najpogostejše oblike modifikacije zgledov so krajšanje oz. izpuščanje nerelevantnih delov, kot so odvisniki ali vrinjeni stavki, poenostavljanje kompleksnih struktur in zamenjava besed ali besednih zvez s pogostejšimi ali ustreznejšimi poimenovanji. Krajšanje se zdi še najmanj sporno in je dejansko velikokrat povsem legitimno z vidika informativnosti, saj stavki večkrat vsebujejo dele, ki so nepotrebni oz. nerelevantni, če niso podani v širšem kontekstu besedila, kot velja recimo za *na primer* v stavku za iztočnico *anonimnost*:

> Jane Austen, na primer, je živela v popolni **anonimnosti**.

Po drugi strani lahko poenostavljanje kompleksnih struktur in zamenjava besed v precej večji meri vplivata na naravnost ali tipičnost zgledov, zato se jima je treba čim bolj izogibati. V nekaterih primerih je sicer zamenjava besed ali besednih zvez

potrebna, npr. pri lastnih imenih, ki jih nadomestimo z zaimki ali generičnimi imeni tipa Janez Novak, ali zaradi izogibanja razžalitvi določene družbene skupine, a tudi tu zadeve niso enoznačne, sploh če gre za javne osebnosti ali če je konkretna oseba tesno povezana z določenim kontekstom rabe iztočnice. Spodnji zgled za *mojstrsko* tako nikakor ne bi imel informativne vrednosti, pa tudi ne bi zvenel naravno, če bi Cristiano Ronaldo zamenjali z generičnim imenom tipa Janez Novak:

> Izid polčasa in tudi končni izid je z **mojstrsko** izvedenim prostim strelom postavil Cristiano Ronaldo.

Pogostost in korenitost spreminjanja zgledov sta odvisni tudi od ciljnega uporabnika slovarja. Pri izdelavi slovarja za nematerne govorce ali mlajše materne govorce, ki še razvijajo jezikovno kompetenco in poznajo manjši nabor besedišča, bo posegov v zglede več,[20] medtem ko lahko pri slovarjih za odrasle materne govorce v zglede posegamo precej manj. Pri odločitvah nas mora voditi tudi namen slovarja: če naj bi uporabnikom pomagal pri jezikovni produkciji in ne samo recepciji, potem je sploh pomembno, da zgledi ostanejo čim bolj naravni in tipični.

Posebna oblika modificiranja zgledov je jezikovno popravljanje. Če najdemo dober korpusni stavek, v katerem manjka vejica, lahko vstavimo vejico in stavek uporabimo v slovarju? Kaj pa, če stavek vsebuje narobe črkovano besedo, besedo v napačnem sklonu ali napačen besedni red? Čeprav se zdi popravek napake črkovanja res malenkosten in včasih nujen poseg, moramo pri popravljanju paziti, saj se lahko hitro znajdemo v drugi skrajnosti, ko vidimo kot napako tudi izbiro ubeseditve ipd. in postane popravljanje zgleda skoraj identično temu, kot če bi si zgled izmislili. Dobro se je držati načela, da skušamo na vsak način najti dober korpusni zgled, ki ne potrebuje popravkov, če pa takšnih zgledov ni na voljo in najdemo potencialno dobre zglede z določenimi (manjšimi) jezikovnimi napakami, jih lahko s popravki vključimo v slovar. Na vsak način pa se je priporočljivo izogniti vnašanju novih napak v slovarske zglede.[21]

5 METODE PRIDOBIVANJA DOBRIH SLOVARSKIH ZGLEDOV

Iskanje (dobrih) slovarskih zgledov je zelo zahteven in potencialno tudi zamuden ter posledično drag proces. Prvič zato, ker je dober slovarski zgled ob upoštevanju vseh v prejšnjem razdelku omenjenih kriterijev zelo težko najti. Drugič, korpusi

20 Dejansko tudi Atkins in Rundell (2008) svoje odobravanje modificiranja zgledov omejita skoraj izključno na slovarje za nematerne govorce.

21 S tega vidika je recimo vprašljiva nerazumljiva odločitev avtorjev SNB, ki so se odločili na koncu (celostavčnih) zgledov opustiti ločila, največkrat piko, kar ob dejstvu, da se zgledi začenjajo z veliko začetnico, ni jezikovno pravilno in tudi na določen način zgledom zmanjšuje avtentičnost.

postajajo vse večji, kar leksikografu sicer ponudi večjo izbiro kandidatov za dobre zglede, a hkrati tudi večjo količino stavkov za analizo. In tretjič, zgledi so ključni element slovarske mikrostrukture, ne samo na ravni gesla, ampak tudi pomenov, podpomenov, kolokacij, stalnih zvez, frazeologije itd. Skratka, leksikograf mora v veliki količini podatkov poiskati veliko dobrih zgledov za vsako slovarsko geslo.

Govorimo lahko o dveh metodah pridobivanja dobrih slovarskih zgledov: ročni in polavtomatski. Pri ročni metodi leksikograf med korpusnimi zadetki oz. konkordancami za določeno iztočnico izbere zglede, pri čemer si lahko pomaga s sortiranjem, filtriranjem in podobnimi funkcijami za urejanje zadetkov. V pomoč mu je lahko tudi vnaprejšnje razvrščanje zgledov glede na kolokacije in slovnične relacije, kar npr. v korpusnem orodju Sketch Engine[22] (Kilgarriff et al. 2004) omogoča funkcija Besedne skice.

Pri polavtomatski metodi leksikografu pomaga orodje za prepoznavo dobrih zgledov, kot je npr. GDEX (*Good Dictionary Example*s; Kilgarriff et al. 2008), ki mu ponudi nabor kandidatov za dobre slovarske zglede, med katerimi leksikograf potem izbere ustrezne. GDEX (gl. tudi razdelek 5.1) razvršča zglede glede na njihovo kakovost pri značilnostih, kot so dolžina zgleda, celostavčna oblika, preprosta ali manj kompleksna skladenjska zgradba povedi, prisotnost ali odsotnost redkih besed, spletnih in elektronskih naslovov ipd. Mnoge od teh značilnosti so posredno povezane s tipičnostjo, informativnostjo in razumljivostjo, torej lastnostmi dobrega zgleda. Značilnosti lahko razdelimo v dve skupini: v prvi so tiste, ki jih zgled mora vsebovati, npr. celostavčnost, odsotnost spletnih naslovov, odsotnost izredno dolgih ali redkih besed ipd. Če zgled ne ustreza vsaj eni od teh značilnosti, dobi toliko kazenskih točk, da se takoj znajde na dnu vseh zadetkov. V drugi skupini so značilnosti, ki so bodisi zaželene bodisi nezaželene (stopnjo (ne)zaželenosti določimo s težo, ki jo pripišemo posamezni značilnosti, in višino dodatnih/odbitih točk), a je pomemben predvsem kumulativni seštevek vrednosti vseh značilnosti v konfiguraciji.

Ključna razlika med metodama je trajanje, saj je polavtomatska metoda precej hitrejša od ročne, a nič manj zanesljiva (Kosem et al. 2012; 2013). V sodobni korpusni leksikografiji tako polavtomatska metoda vse pogosteje nadomešča ročno, zlasti pri projektih, kjer se izdelujejo slovarske baze, ki vsebujejo več zgledov kot na njih temelječi slovarji (npr. nizozemski ANW).

6 ZGLEDI V SLOVARJU SODOBNEGA SLOVENSKEGA JEZIKA

V tem razdelku se posvetimo vlogi zgledov v predlaganem slovarju sodobnega slovenskega jezika, in sicer načinu njihovega pridobivanja in beleženja, razliki

22 http://www.sketchengine.co.uk/ (dostop 8. 8. 2015).

med zgledi v slovarski bazi in slovarju, nekaj besed pa namenimo tudi različnim možnostim vizualizacije zgledov, ki jih ponuja digitalni medij.

6.1 Način pridobivanja in beleženja slovarskih zgledov

Pridobivanje zgledov je, prek orodja GDEX, sestavni del različice polavtomatske metode, imenovane avtomatsko luščenje leksikalnih podatkov (ALLP; Kosem et al. 2012), pri kateri se podatki (skladenjske strukture, kolokacije in zgledi, pa tudi določene informacije o iztočnici in predlogi za slovnične oznake) prek orodja Sketch Engine oz. aplikacije Besedne skice z uporabo API-skripte (angl. *Application Programming Interface*) avtomatsko izvozijo iz korpusa neposredno v program za izdelavo slovarjev, kjer jih leksikograf pregleda, selekcionira in uredi.[23] Metoda leksikografu še vedno ponudi dovolj podatkov za temeljito analizo in izdelavo gesla. Izkušnje pri izdelavi LBS kažejo, da z uporabo takšne metode leksikograf pri izdelavi gesla ne pregleda nič manj zgledov (pogosto jih celo več!), kot bi jih z uporabo ročne ali polavtomatske metode znotraj korpusnega orodja. Prednosti metode so odprava nepotrebnega kopiranja podatkov iz korpusnega orodja in njihovega vnašanja v program za izdelavo slovarjev ter posledično hitrejša, pa tudi bolj razpršena in s tem zanesljivejša analiza.

Za zglede ključni del ALLP je priprava ustrezne konfiguracije oz. konfiguracij za orodje GDEX. Že od leta 2011 je na voljo različica konfiguracije GDEX za slovenščino (Kosem et al. 2011), ki smo jo izdelali za namene izdelave LBS in je dokaj uspešno dosegala cilj vsaj treh dobrih zgledov od desetih ponujenih na kolokator, kar je bila privzeta nastavitev pri izdelavi LBS. Vendar pa omenjena različica ne zadosti potrebam ALLP, kjer izvažamo **prvih** X (največkrat od tri do pet) ponujenih zgledov na kolokator in morajo biti dejansko vsi potencialno dobri. Poleg tega so se že pri uporabi prvotne različice pri analizi v orodju Sketch Engine za potrebe izdelave LBS opazile precejšnje razlike v kvaliteti ponujenih zgledov po posameznih besednih vrstah. Zato smo pri testiranju ALLP pri izdelavi LBS za vsako besedno vrsto izdelali samostojno konfiguracijo GDEX. Izdelava konfiguracij je potekala v dveh korakih: najprej smo na podlagi analize dobrih zgledov v LBS oblikovali izhodiščno konfiguracijo za vsako besedno vrsto, potem pa smo s prilagajanjem nastavitev izdelali nove različice, katerih evalvacija je vedno potekala tako, da smo rezultate nove različice primerjali z rezultati prejšnje. Postopek smo ponavljali, dokler nismo izoblikovali optimalne končne verzije konfiguracije GDEX za postopek ALLP. Pomemben rezultat tega dela analize je oblikovanje več novih klasifikatorjev, ki jih prvotna verzija GDEX ni vključevala. Klasifikatorji so tako sledeči:

23 O podobni metodi sta razmišljala že Rundell in Kilgarriff (2011).

- Cela poved. Na ta način prioritiziramo zglede, ki ustrezaju načelu celostavčnosti.
- Ne vsebuje pojavnic s frekvenco manj kot 3. Iščemo zglede, ki ne vsebujejo zelo redkih besed, napak in korpusnega šuma.
- Poved mora biti daljša od 7 pojavnic. Nočemo prekratkih zgledov, saj jim velikokrat manjka kontekst. Lažje je krajšati daljše zglede kot iskati nove.
- Poved mora biti krajša od 60 pojavnic. Izločamo samo res dolge povedi, daljše povedi vedno lahko okrajšamo.
- Poved ne sme vsebovati ponovitve iztočnice. Gre za pomemben klasifikator, kajti večkratno ponavljanje iztočnice zgledu jemlje razumljivost in informativnost.
- Vsebuje elektronski ali spletni naslov. Zgledi, ki ustrezajo temu kriteriju, prejmejo visok kazenski pribitek.
- Optimalna dolžina (med X in Y pojavnic). Medtem ko s klasifikatorjema za minimalno in maksimalno dolžino povedi na nek način izločamo prekratke in predolge povedi (jih potiskamo na dno seznama), z optimalno dolžino nagrajujemo povedi z dolžino znotraj danega razpona. Najpogosteje je optimalna dolžina zgleda med 15 in 40 pojavnic, odvisno od besedne vrste. Analiza dobrih zgledov v LBS pri izdelavi prvotne različice (Kosem 2012) je npr. pokazala, da je povprečna dolžina zgledov za pridevniške iztočnice 28,64 pojavnice, za prislove 27,03 pojavnice in za prislov 27,39 pojavnice.
- Vsebuje redke leme. Klasifikator dodeli točkovni odbitek povedi za vsako redko lemo, ki jo vsebuje. Frekvenčna meja, ki opredeljuje redkost, je odvisna od velikosti korpusa.
- Vsebuje pojavnice, daljše od 12 znakov. Klasifikator kaznuje vsako pojavnico, ki izpolnjuje omenjeni kriterij. Analiza je namreč pokazala, da so daljše pojavnice največkrat nebesede ali korpusni šum.
- Število ločil v zgledu (brez vejic). Klasifikator točkovno kaznuje poved, v kolikor je preseženo določeno število ločil v njej, pri čemer se vejice ne upoštevajo.
- Število vejic v povedi. Klasifikator točkovno kaznuje povedi z več kot tremi vejicami, saj je bilo ugotovljeno, da so takšne povedi pogosto kompleksnejše in posledično slabši kandidati za dobre zglede.
- Pojavnice z velikimi začetnicami. Klasifikator točkovno kaznuje povedi, ki vsebujejo pojavnice z velikimi začetnicami, in je namenjen predvsem kot dopolnilo klasifikatorju za lastna imena.

- Pojavnice z mešanimi simboli (npr. črke in številke). Klasifikator točkovno kaznuje nebesede in korpusni šum.
- Lastna imena. Klasifikator točkovno kaznuje povedi s pojavnicami, ki so v korpusu označene kot lastna imena. Če je takšnih pojavnic v povedi več, je kaznovana vsaka posamezna pojavitev.
- Zaimki. Klasifikator z odbitkom kaznuje vsako pojavitev zaimka v povedi. Klasifikator je koristen predvsem, ko je zaimkov v povedi več, saj so takšne povedi ponavadi manj razumljive oz. potrebujejo dodaten kontekst.
- Položaj leme v povedi. Klasifikator točkovno kaznuje povedi, kjer se lema pojavlja izven določenega razpona v povedi. Tako je bilo za glagolske leme ugotovljeno, da so boljši kandidati za dobre zglede tiste povedi, v katerih se glagol ne pojavlja na začetku, tj. v prvih 40 ostotkih pojavnicah povedi.
- Seznam prepovedanih besed na začetku povedi. Pri izdelavi konfiguracij se je izkazalo, da so določene besede na začetku povedi že dober indikator, da ne gre za dobrega kandidata za slovarski zgled. Gre za besede, kot so *sledi, tovrsten, oboji* ipd., ki nakazujejo, da poved zahteva dodaten kontekst. Za namene klasifikatorja je bil na podlagi analize in opažanj pri evalvaciji konfiguracij izdelan seznam takšnih besed. Klasifikator tako točkovno kaznuje povedi, ki se začenjajo s katero od besed na seznamu.
- Seznam prepovedanih besednih zvez na začetku povedi. Podoben klasifikator kot klasifikator za prepovedane besede, s tem da kaznuje pojavitev določenih večbesednih nizov na začetku povedi.
- Drugi kolokator. Eden najpomembnejših klasifikatorjev, ki točkovno nagrajuje zglede, ki vsebujejo najbolj tipične kolokatorje določene kolokacije, in s tem posredno upošteva merilo koligacijske tipičnosti. Npr. pri kolokaciji *klavrn* + *podoba* klasifikator dodeli dodatne točke zgledom s statistično pomembnim drugim kolokatorjem *kazati*, pri čemer se izkaže, da tako identificirani zgledi pa vsebujejo tudi tipično širšo strukturo kolokabilne okolice: *kazati klavrno podobo česa*.
- Levenshteinova razdalja. Gre za algoritem,[24] ki meri podobnost med nizi, v našem primeru povedmi. Če klasifikator najde dve podobni ali celo enaki povedi, tisto z nižjim točkovanjem vrže na dno seznama zadetkov.

Večina razlik med konfiguracijami za različne besedne vrste se pojavlja v nastavitvah posameznih klasifikatorjev, čeprav najdemo tudi razlike med klasifikatorji (npr. samo konfiguracija za glagol vsebuje dodaten klasifikator položaja leme v

24 http://en.wikipedia.org/wiki/Levenshtein_distance (dostop 8. 8. 2015).

povedi). Vsaki povedi je pripisana določena vrednost med 0 in 1, ki predstavlja seštevek vseh vrednosti klasifikatorjev (pri čemer vsakemu klasifikatorju določimo določeno težo pri končnem izračunu) in služi kot osnova orodju GDEX pri razvrščanju kandidatov za dobre zglede za vsako kolokacijo ter posledično odloča o zgledih, ki jih izvozimo z metodo ALLP.

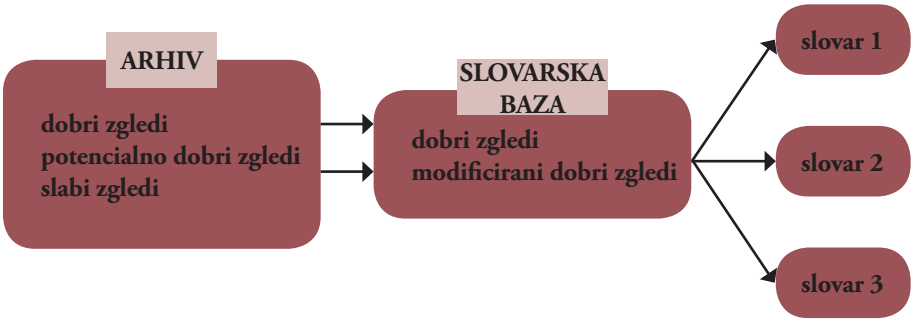
Za vsak izluščeni zgled je koristno izvoziti tudi metapodatke o besedilu, iz katerega zgled izvira, npr. leto, vir, avtor, naslov ipd. To nam zagotavlja sledljivost zgleda, ponuja pa tudi možnost prikaza tovrstnih informacij v slovarju. Kot pri vseh ostalih delih slovarja se namreč nikoli ni dobro omejevati samo na potrebe konkretnega slovarja, saj je kasnejše iskanje informacij o zgledih v korpusu precej zamudnejše in dražje. Dober primer koristi metapodatkov je povezan s posodabljanjem slovarja: če namreč želimo enkrat zamenjati stare zglede s sodobnejšimi, s pomočjo podatka o letu nastanka besedila lahko takoj poiščemo vse zglede, ki so nastali pred določenim letom. Metapodatki so lahko koristni tudi pri preprečevanju vključevanja izrazito ideoloških zgledov v slovarska gesla. Pri zgledih, ki jih izluščimo z orodjem GDEX, smo lahko pozorni na primere, ko je večina zgledov v geslu oz. pri posameznem pomenu iz enega samega vira ali samo nekaj virov (prim. analizo zgledov za *pederastija* v Gorjanc 2014).

6.2 Zgledi v slovarski bazi in slovarski zgledi

Pri diskusiji o iskanju in beleženju zgledov je treba upoštevati tudi vlogo odnosa med slovarjem in slovarsko bazo, pa tudi obeh virov do arhiviranih podatkov (Slika 1). Postopki, opisani v tem prispevku, so najbolj aktualni za izdelavo slovarja sodobnega slovenskega jezika, ker pa se bo jezikovni opis delal povsem na novo, bo marsikateri del podatkov (tudi zgledi), zabeleženih pri analizi korpusa oz. korpusov, uporaben pri izdelavi drugih slovarjev. Iz enega zgleda lahko izpeljemo več različic za različne uporabnike oz. slovarje; za odrasle materne govorce lahko določen dober zgled ostane nespremenjen, medtem ko ga za mlajše govorce po potrebi lahko malenkostno spremenimo, npr. ga skrajšamo ali nadomestimo redko besedišče s pogostejšim, ki ga mlajši govorci poznajo. Zaradi takšne večnamenske rabe zgledov moramo izvirno izluščene zglede in vse z njimi povezane metapodatke arhivirati.

Arhiv izvirnih zgledov omogoča tudi analizo modifikacij zgledov in njihov obseg ter posledično izboljšavo konfiguracij za njihovo luščenje. Tudi slabe oz. nerelevantne zglede, ki so sestavni del avtomatsko izluščenih podatkov in jih v bazi nočemo, je treba arhivirati, saj njihova analiza lahko razkrije nove lastnosti slabih zgledov, ki jih potem integriramo v konfiguracije za luščenje zgledov. Podoben postopek je bil že uporabljen pri izdelavi prve različice konfiguracije GDEX za

slovenščino (Kosem et al. 2011), kjer so se parametri klasifikatorjev testne konfiguracije, izdelane na podlagi analize obstoječih, ročno izbranih zgledov v LBS, izboljšali na podlagi analize izbranih (dobrih) in neizbranih (slabih) zgledov pri pregledovanju testnih rezultatov. Poleg tega ne smemo pozabiti na vlogo slovarskih podatkov pri izdelavi ostalih jezikovnih tehnologij za slovenščino. Skratka, že v samem načrtovanju slovarja je treba posvetiti veliko pozornosti tipom podatkov v slovarski bazi in njihovemu beleženju ter dejansko jemati slovar, tudi splošnega, kot eno izmed njenih izpeljank.



Slika 1: Zgledi z vidika razmerja med arhivom, slovarsko bazo in slovarji.

Ker slovarska baza vsebuje več podatkov kot sam slovar, se poraja vprašanje, koliko dodatnega časa vzamejo leksikografski ekipi, ki je osredotočena na izdelavo slovarja, vsi postopki beleženja dodatnih podatkov. Zato mora načrtovanje slovarske baze slediti dvema načeloma: avtomatizirati čim več (rutinskih) leksikografskih postopkov in poskrbeti, da nobena leksikografska odločitev ne ostane neizkoriščena. S tega vidika je uporaba metod, kot je ALLP, pravzaprav nujna, saj si drugače ne moremo predstavljati izdelave baze in slovarja v roku, ki bi zadovoljil tako uporabnike kot financerje, kar ugotavljajo tudi pri podobnih leksikografskih projektih v tujini (npr. nizozemski slovar ANW). Za primer lahko vzamemo povsem osnovno nalogo: vtipkavanje iztočnice in njene besedne vrste v geselski članek v slovarski bazi. Recimo, da za vpis teh dveh podatkov porabimo povprečno 5 sekund. Pri 100.000 iztočnicah to znese 500.000 sekund oz. malo manj kot 139 ur. Pri ALLP se ta dva podatka pripišeta avtomatično – torej prihranimo skoraj en človek-mesec. Precej podobno velja za leksikografske odločitve. Pri ročni analizi oz. analizi v korpusnem orodju, tudi z uporabo orodja GDEX, leksikograf pregleda veliko zgledov in pri vsakem sprejme načelno odločitev, ali je dober ali ne. Ker pa leksikograf iz korpusnega orodja v bazo prenese samo dobre oz. kandidate za dobre zglede, so arhivirane samo tovrstne odločitve. Pri metodi ALLP lahko v bazi beležimo čisto vsako odločitev: identifikacijo dobrega zgleda (zgled ostane v bazi v nespremenjeni obliki), potencialno dobrega zgleda (zgled je modificiran) in slabega zgleda (brisanje zgleda).

V postopek izbire zgledov lahko vpeljemo tudi množičenje in tako še dodatno razbremenimo leksikografsko ekipo. A ker mora dober slovarski zgled vsebovati kombinacijo različnih lastnosti, ki se jih mora leksikograf naučiti prepoznati, si je težko predstavljati, da bi takšno delo lahko prepustili neleksikografom oz. procesu množičenja. Da to postane izvedljivo, se moramo zavedati značilnosti in omejitev množičenja (gl. Čibej et al. 2015a in Fišer in Čibej 2015). Kot prvo, naloge morajo biti enostavne, množičarji izbirajo samo med odgovori *Da*, *Ne* in *Ne vem*. Poleg tega naloge ne smejo temeljiti na določanju težko določljivih elementov (kot so npr. lastnosti dobrega zgleda) ali stopenjskosti; vprašanja tipa *Ali je to dober slovarski zgled?* in *Kako dober je ta zgled?* torej ne pridejo v poštev. Kot so pokazala testiranja množičenja na zgledih v LBS (Kosem et al. 2013), je najbolj smiselno uporabiti zglede pri odkrivanju napačnih podatkov (npr. ko raba iztočnice in kolokatorja v zgledu ne ustreza skladenjski strukturi, v kateri naj bi se kolokacija pojavljala) ali razporejanju kolokacij in z njimi povezanih zgledov pod relevantne pomene in podpomene.

6.3 Vizualizacija zgledov

Leksikografsko delo se z vidika zgledov ne konča z zabeleženjem dobrih zgledov, saj ima svojo vlogo pri uspešnosti izpolnjevanja namena zgledov tudi način predstavitve zgledov slovarskemu uporabniku. Zato je ključno sodelovanje med leksikografi in oblikovalci oz. ekipo, ki skrbi za vizualizacijo slovarskih podatkov. Še tako dober zgled je precej odvisen od ustrezne vizualizacije, kot je na primer izbira ustrezne pisave, ki vpliva na berljivost besedila in tudi na to, kako dobro si uporabnik zapomni prebrano informacijo (Nesi 2011). Če upoštevamo, da zgledi predstavljajo precej velik, če ne celo največji, odstotek besedila v slovarju, je izbira načina vizualizacije še kako pomembna.

Uporabniku lahko pri branju zgledov pomagamo tudi z označevanjem iztočnice, njenih stalnih zvez ali podobnih delov gesla. Sploh v sodobni leksikografiji, ko se prikazujejo celostavčni, daljši zgledi, se zdi smiselno uporabnika opozoriti na iztočnico in s tem na del zgleda, ki naj bi mu posvetil večjo pozornost. Največkrat se za tako označevanje uporabi krepki tisk, v elektronskih slovarjih najdemo tudi rabo drugačne barve (Slika 2), precej redkeje pa ležeči tisk, kar je tudi posledica dejstva, da je ležeči tisk oblika pisave, ki jo za prikazovanje zgledov uporablja večina slovarjev. Slednja možnost se zdi tudi manj učinkovita (Slika 3). Druga vloga krepkega tiska oz. uporabe pisave, ki se razlikuje od pisave ostalega dela zgleda, je izpostavitev določenih zelo tipičnih kolokacij, stalnih zvez in fraz (Slika 4). Odločitev o uporabi oz. obsegu in načinu označevanja iztočnice in ostalih delov zgleda mora biti v vsakem primeru podprta z uporabniško študijo.

fach

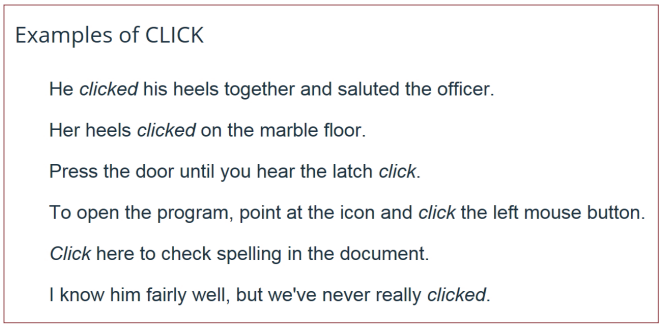
Prawie 60 lat temu zaczął się uczyć fryzjerskiego fachu i nadal pracuje w zawodzie.

źródło: NKJP: Katarzyna Skrzypek: Cyrkiel za uszami, Dziennik Zachodni, 2005-03-31

Miał dobry fach - przez kilkanaście lat pracował w dużej warszawskiej fabryce jako spawacz, był też ślusarzem, szlifierzem i monterem.

źródło: NKJP: Monika Mikołajczuk: Między Kantem a Wolterem, Polityka, 2001-07-14

Slika 2: Rdeče obarvana iztočnica v zgledih v Velikem slovarju poljskega jezika.



Examples of CLICK

He *clicked* his heels together and saluted the officer.

Her heels *clicked* on the marble floor.

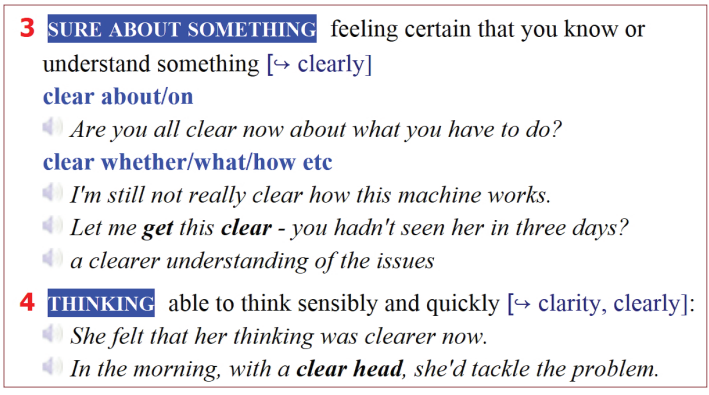
Press the door until you hear the latch *click*.

To open the program, point at the icon and *click* the left mouse button.

Click here to check spelling in the document.

I know him fairly well, but we've never really *clicked*.

Slika 3: Iztočnica v ležečem tisku v zgledih v slovarju Merriam-Webster.



3 SURE ABOUT SOMETHING feeling certain that you know or understand something [↪ clearly]

clear about/on

Are you all clear now about what you have to do?

clear whether/what/how etc

I'm still not really clear how this machine works.

*Let me **get** this **clear** - you hadn't seen her in three days?*

a clearer understanding of the issues

4 THINKING able to think sensibly and quickly [↪ clarity, clearly]:

She felt that her thinking was clearer now.

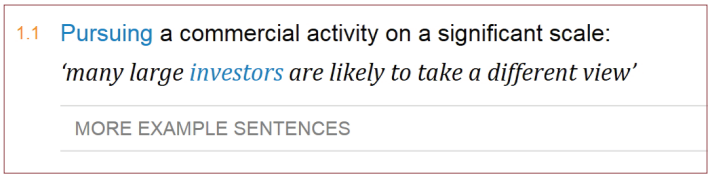
*In the morning, with a **clear head**, she'd tackle the problem.*

Slika 4: Izpostavljene kolokacije in fraze v zgledih v slovarju Longman.[25]

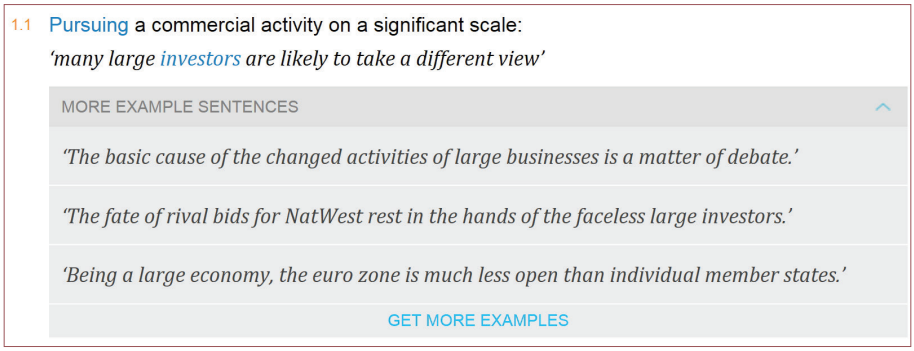
25 http://www.ldoceonline.com/ (dostop 8. 8. 2015).

Omenili smo že, da je o vsakem zgledu v slovarski bazi dobro imeti čim več metapodatkov, in tudi takšne podatke lahko ponudimo uporabniku. Prikazovanje tovrstnih podatkov, zlasti v splošnih enojezičnih slovarjih, je sicer redkost – izjema je npr. Veliki slovar poljskega jezika (Slika 2) –, iz preprostega razloga: metapodatki, kot so vir, avtor in naslov, imajo referenčno vlogo in sugerirajo, da je bil zgled vzet neposredno iz določenega vira, kar leksikografom odvzame možnost vnašanja kakršnih koli sprememb. Drugi razlog proti prikazovanju metapodatkov je, da predstavljajo za uporabnike dokaj nenujne podatke, ki jemljejo dragoceni prostor na zaslonu in največkrat odvračajo njihovo pozornost od osrednje vloge zgleda, namreč prikazovanja rabe iztočnice v določenem (pod)pomenu.

Načelo informativnosti slovarskih zgledov med drugim leksikografom narekuje omejevanje glede števila ponujenih zgledov na geslo ali posamezen pomen v njem. Pa vendar se hitro zgodi, da imamo večje število zgledov na (pod)pomen, skladenjsko strukturo ali kolokacijo, kar lahko postane vizualizacijska težava zlasti pri pomensko zelo razčlenjenih geslih. Ena od dobrih rešitev je privzeti prikaz samo določenega števila zgledov, medtem ko se dodatni zgledi uporabniku razkrijejo po kliku na gumb (Slika 5 in Slika 6). Vse več spletnih slovarjev ponuja tudi neposredno povezavo na korpusne zadetke, kar je za zahtevnejše uporabnike vsekakor dobrodošla funkcija, saj si lahko ogledajo številne primere realne rabe iztočnice, njene kolokacije, stalne zveze ipd.



Slika 5: Povezava za prikaz dodatnih zgledov (More example sentences) v slovarju Oxford.



Slika 6: Prikazani dodatni zgledi v slovarju Oxford.

7 ZAKLJUČEK

Zaradi svoje nepogrešljivosti v slovarskem geslu na eni strani in zahtevnega načina pridobivanja na drugi bomo zgledom pri snovanju slovarja sodobnega slovenskega jezika posvetili veliko pozornosti. V navodilih leksikografom bomo jasno opredelili kriterije dobrih zgledov skupaj s konkretnimi primeri dobre in slabe prakse, pa tudi z vidika ideološkosti, ter izdelali oz. uporabili orodja, ki zagotavljajo čim večjo doslednost pri upoštevanju teh kriterijev. Hkrati bomo pripravili primere dovoljenih modifikacij in posledično tudi ustrezen sistem arhiviranja izvirno izluščenih zgledov. Ker naj bi SSSJ predstavljal pomemben vir tudi jezikovnotehnološki skupnosti, je nujno zagotoviti čim večje število zgledov in jih opremiti s čim več metapodatki.

Ravno zaradi teženj po vključevanju čim večjega števila zgledov v slovar oz. slovarsko bazo je nujna uporaba polavtomatskih ali avtomatskih metod pri pridobivanju zgledov. V nasprotnem primeru se namreč izdelava slovarja lahko zavleče do takšne mere, da so zgledi že pred dokončanjem slovarja potrebni zamenjave. Iz takšnega razmišljanja izhaja metoda ALLP, ki jo predlagamo za iskanje in beleženje zgledov v slovarju sodobnega slovenskega jezika ter predstavlja nov način leksikografske analize in iskanja zgledov ne samo v slovenskem, ampak tudi v mednarodnem prostoru. Na podlagi slovenskih izkušenj, pridobljenih pri izdelavi LBS, so podobno metodo že začeli uporabljati v Estoniji, in sicer pri izdelavi kolokacijskega slovarja za nematerne govorce estonščine (Kallas et al. 2015).

Pomembna naloga leksikografske skupnosti v zvezi z zgledi pa je tudi izvajanje (nadaljnjih) raziskav o tem, kako, kdaj in na kakšen način slovarski uporabniki uporabljajo zglede ter kakšni zgledi uporabnikom najbolj koristijo. Na podlagi izsledkov bo mogoče v prihodnje še izboljšati postopke izbiranja zgledov, kot tudi način in količino njihovega podajanja v slovarjih.

Homonimija in večpomenskost: od teorije do slovarja

Polona Gantar

Abstract
This paper discusses homonymy and polysemy from a theoretical perspective as well as their practical role in dictionaries. Specifically, the focus is on the approaches and efficiency of their solutions in different dictionaries for various users. Firstly, different problematic cases, as discussed by Atkins and Rundell (2008: 192-193), are outlined, and then a treatment is presented of homonymy and polysemy in selected English monolingual dictionaries and existing Slovenian dictionaries, with a focus on their key advantages and shortcomings. In the last part of the paper, a proposal is presented for the treatment of homonymy and polysemy in the proposed dictionary of contemporary Slovenian, taking into account the best practices in state-of-the-art lexicography, the needs of potential users, and the fact that the dictionary will be (primarily) designed for the online format.

Keywords: homonymy, polysemy, user oriented dictionary, web dictionary

Ključne besede: homonimija, večpomenskost, uporabniška naravnanost slovarja, spletni slovar

1 UVOD

Homonimija[1] in večpomenskost sta v jezikoslovju pogosto obravnavana semantična problema, njuna obravnava v kontekstu pričujoče monografije pa je namenjena prikazu možnosti obravnave, ki izhaja iz spletne zasnove slovarja in uporabniške naravnanosti, ne da bi se (oz. prav zato, da se ne bi) pri tem izgubljala tudi kakovost slovarskega opisa.

V zvezi s homonimijo se običajno sprašujemo, kaj je dejansko to, kar imajo določene besede skupnega? Kako lahko iste (oz. enake) besede pomenijo različne stvari? Problema, ki je v izhodišču semantične narave, se zavedajo predvsem leksikografi, saj prideta semantična bližina in referenčna razpršenost še posebno do izraza, ko so besede obravnavane zunaj svojega naravnega okolja – besedilnega konteksta (Frath 2000: 44; Atkins in Rundell 2008: 269). V vsakdanjih govornih situacijah namreč raba besed skoraj nikoli ni dvoumna. Drugačno je stanje v slovarjih, kjer je večpomenskost načeloma obravnavana znotraj ene iztočnice, homonimija pa znotraj več samostojnih iztočnic, vendar pa odločitve in strategije prikazovanja homonimije v odnosu do večpomenskosti v slovarjih niso enotne, kar je jasna indikacija tega, da razmerje ni jasno razmejljivo in absolutno.

Namen našega prispevka zato ni prikazati določen način obravnave homonimije kot edino ustrezen, ampak utemeljiti prikaz izrazne podobe leme v odnosu do njene pomenske razpršenosti na način, ki je primarno usmerjen v reševanje potencialnih uporabniških zadreg in ki upošteva lastnosti spletnega medija, manj pa v iskanje možnosti čim bolj doslednega prenosa teoretičnih predpostavk v slovarski priročnik.

2 HOMONIMIJA KOT TEORETIČNI PROBLEM

Obravnava homonimije v povezavi z večpomenskostjo predstavlja z združitvijo teoretičnega in praktičnega vidika, kot pravi Taylor (2003), t. i. »polisemični paradoks«, saj govorci v vsakdanjih sporočanjskih situacijah rešujejo homonimijo brez težav, medtem ko teoretični pristopi naletijo na probleme tam, kjer skušajo ločevati med večpomenskostjo in homonimijo in zagotoviti jasno določen opis leksikalne dvoumnosti. Kot pravi Frath (2000: 44), je leksikografska obravnava, ki temelji na takem izhodišču, nujno problematična, saj slovar ni namenjen pojasnjevanju delovanja jezika, ampak predvsem zagotoviti »posnetek jezikovne rabe v določenem trenutku«.

1 Izraz homonimija se sloveni kot *enakoizraznost* in vključuje tako *enakozvočnice* kot *enakopisnice* (Vidovič Muha 1997: 9). V prispevku uporabljamo mednarodno uveljavljen izraz v najširšem pomenu.

2.1 Semantična izhodišča

Znotraj semantičnih opredelitev sta se uveljavila dva temeljna principa ugotavljanja večpomenskosti, ki posledično določata tudi obravnavo homonimije. Prvi temelji na razmerju dobesedni in preneseni pomen (t. i. linearna teorija, Frath 2000: 44) ter na prepoznavanju metonimičnih in metaforičnih prenosov ter polisemičnih regularnosti (prim. Apresjan 1973; 2002; za slovenščino Vidovič Muha 2000; Snoj 2010), drugi pa na ugotavljanju razmerja med splošnim in specializiranimi pomeni v sobesedilu.

2.1.1 Dobesedni in preneseni pomen

Prvi pristop predvideva, da obstaja razmerje med dobesednim in prenesenim, tj. iz njega izpeljanim pomenom. Npr. dobesedni pomen[2] besede *posteljica* v slovenščini je 'kos pohištva', lahko pa pomeni tudi 'organ, v katerem se razvija in rodi otrok' (*iztisniti posteljico*), 'podlago, na kateri kaj leži' (*posteljica iz rukole*) itd., kar je mogoče obravnavati kot prenesene pomene. Vendar pa dobesednega pomena ni vedno mogoče jasno določiti. Naprimer beseda *položaj* lahko pomeni 'fizično držo' (*sedeč položaj*), 'psihično stanje' (*brezizhoden položaj*), 'sociološko opredelitev' (*enakopraven, privilegiran položaj*) ali 'prostorsko umestitev' (*izpostavljen položaj*), pri čemer nobeden od pomenov, razen dogovorno, ni sam po sebi bolj dobeseden kot drugi.

2.1.2 Splošni in specializirani pomen

Drugi način obravnave večpomenskosti predvideva, da ima vsaka beseda nek splošni pomen, medtem ko se specializirani pomeni ustvarjajo v vsakokratnem besedilnem okolju. Hanks (2013) v zvezi s tem pravi, da besede same na sebi nimajo pomena, pač pa le tendence, da bi nekaj pomenile, te tendence pa se lahko realizirajo zgolj v (so)besedilu. Vendar pa, čeprav se zdi ustrezno razpoznavanje pomena besed na podlagi sobesedila neproblematično v večini primerov, nekateri avtorji opozarjajo, da sporočanjski proces v situaciji, kjer je razumevanje vsake besede odvisno od razumevanja druge, pravzaprav sploh ne more začeti potekati (Frath 2000: 45). Pustejovsky (1995) je tako v okviru generativnega pristopa pri kontekstualnem prepoznavanju pomenov skušal s pomočjo pretvorbenih pravil določiti štiri nivoje prepoznavanja leksikalnega védenja iz jezikoslovno definiranih izrazov, tj. njihovo leksikalno, udeleženo, dogodkovno in t. i. »qualia«

2 Pomeni, ki jih navajamo v prispevku, so, če ni povedano drugače, določeni na podlagi analize v korpusu Gigafida in se ne ujemajo nujno s stanjem v obstoječih priročnikih.

zgradbo, pri čemer je zadnja določena še s formalno, konstitutivno, dogodkovno in agentivno vlogo. Vendar pa tak pristop v praksi ne deluje vedno, saj se pri večpomenskih besedah za ustrezno interpretacijo na podlagi sobesedila ponuja več možnih vlog. Tako se lahko po vzoru Pustejovskega[3] vprašamo, kako ustrezno pomensko razdvoumiti zvezo *začeti s kosilom* v hipotetičnem stavku *Najbolj pripravno je začeti s kosilom*, če imamo na voljo več interpretacij, ki so povezane s pomensko različnimi glagoli, ki se sicer sopojavljajo z elementom *kosilo*, npr. *začeti kuhati kosilo, začeti jesti kosilo, iti na kosilo* itd., hkrati pa imajo ti pomensko različni glagoli določene sobesedilne elemente prekrivne, npr. *iti na/skuhati/pripraviti/jesti … kosilo/testenine/špagete.*

Zdi se, da nesporazum glede uspešnosti razbiranja ustreznega pomena iz sobesedila nastane takrat, ko želimo ustvariti dovolj umetne situacije, ki bi pokrile čim več možnosti, ki jih vključujejo teoretična predvidevanja.

2.1.3 Pomensko razdvoumljanje besed na podlagi korpusa

Kot zatrjujeta Atkins in Rundell (2008: 269), prisotnost naravnega sobesedila vedno omogoči izbiro ustreznega pomena, zato v naravnih sporočanjskih situacijah proces pomenskega razdvoumljanja skoraj nikoli ni neuspešen. Količina (so)besedila, ki je potrebna za ustrezno pomensko razdvoumljanje, pa je različna in sega od neposrednih kolokacij do besedilnih tipov, žanrov in zunajjezikovnih elementov, ki sooblikujejo vsako sporočano besedilo. V potrditev teze, da je v realnih situacijah sobesedilo navadno zadosten razreševalec pomenske dvoumnosti, lahko navedemo naš zgornji primer v celoti: *Najbolj pripravno je začeti s kosilom ali večerjo, in če se bosta imela lepo, lahko druženje nadaljujeta s sprehodom po mestu,* kjer elementi sobesedila in prepoznavna sporočanjska situacija pripomorejo k interpretaciji zveze bolj v smeri pomena 'iti na' kosilo, kot pa npr. 'začeti kuhati' ali 'pripravljati kosilo'.

Na sobesedilu kot ključnem elementu pomenskega razdvoumljanja temelji tudi Hanksova teorija jezikovnih konvencij ter možnosti njihove izrabe (angl. *Theory of Norms and Exploitations*, Hanks 1994; 2013; Hanks in Pustejovsky 2004; 2005), ki sistematično ločuje med pravili, ki določajo konvencionalno jezikovno rabo, in pravili, ki omogočajo ustvarjalno jezikovno rabo. Hanks na številnih korpusnih primerih pokaže, da ima vsakokratna jezikovna raba določenega pomena v prepoznavnem sobesedilnem vzorcu ključno vlogo pri tem, kako ljudje razbiramo pomen iz sobesedila. Analiza tudi pokaže, da so številni pomeni enkratni in kot taki slovarsko nezanimivi, hkrati pa z vidika udeležencev v

3 Pustejovsky (1995: 32) je za prikaz zgornjega problema uporabil primer *Mary began the novel*, glagol *began* – 'je začela' – v kombinaciji z *book* – 'knjiga' – pa naj bi govorec razumel v pomenu 'pisati' in ne npr. 'brati'.

sporočanjski situaciji nikoli niso problematični v smislu ustreznega razumevanja. Ker torej temelji analiza korpusnih vzorcev podobno kot FrameNet (Fillmore et al. 2004) na sobesedilni indikaciji pomenov, sicer zastopanih z določeno izrazno enoto, je problem homonimije kot popolne pomenske različnosti dveh enakih izraznih enot vedno obravnavan kot večpomenskost. Podobno tudi pomenske sheme znotraj FrameNeta vključujejo številne leme, pri čemer se ista lema lahko pojavlja v več različnih shemah, pojavitev v določeni shemi pa je pogojena z njenimi skladenjskimi in pomenskimi lastnostmi, ki jih shema določa, kot so npr. obvezni jedrni elementi ipd., efektivno pa ne ločuje med načini izražanja pomenskih razmerij v smislu večpomenskosti na eni ter homonimije na drugi strani (ibid.: 1091).

2.2 Razumevanje leksema kot dvočlenskega jezikovnega znaka

Na izhodiščih linearne teorije temelji tudi ločevanje homonimije od večpomenskosti v leksikalni teoriji,[4] ki pojmuje leksem kot dvočlenski jezikovni znak, v katerem je vsebina[5] neločljivo povezana z njegovim izrazom. Pojav homonimije je tako obravnavan kot prekrivnost materialne jezikovne danosti, tj. kot razmerje na izrazni ravni, ob popolnoma različnem pomenu oz. celotnem pomenju, ki se obravnava kot razmerje na pomenski ravni. S pojmom popolne različnosti vsebine pa je mišljena različnost na ravni pomenskih sestavin v okviru metaslovarsko določene pomenske razlage, ki naj bi odražala tudi vsebinske predstavne lastnosti celotne jezikovne skupnosti (Vidovič Muha 2000: 179).

Tak vidik združuje sinhroni pristop, kjer se skuša prepoznati trenutno jezikovno vedenje določene jezikovne skupnosti (Vidovič Muha 1997: 11, 13), in diahroni pristop, kjer se išče posameznemu homonimu njegov izvor prek zgodovinskorazvojne dinamike. Če s sočasnega vidika velja za homonimni niz popolna izrazna prekrivnost in hkrati popolna vsebinska neprekrivnost, se moramo torej zanašati na predpostavko, da bo tako razlikovanje nedvoumno prepoznano tudi pri vseh ali vsaj večini govorcev. Kot smo opisali zgoraj, pa so prav ta razmerja nepredvidljiva in podrejena individualnim interpretacijam. Pri prenosu te teorije v slovarsko prakso se tako kaže, da so odločitve večkrat prepuščene leksikografovi sprotni presoji in ne veljajo nujno za celotno jezikovno skupnost.

4 Posledično pa tudi v slovarjih, ki temeljijo na teoriji leksikalnega pomena, kamor sodijo Slovar slovenskega knjižnega jezika (SSKJ) in njegova druga izdaja (SSKJ2), Slovar novejšega besedja slovenskega jezika (SNB) in Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika (NSSKJ).

5 Po Vidovič Muha (2000: 18) je vsebina v slovarju aktualizirana kot pomen, sicer pa opredeljena tudi kot virtualna, pojmovna danost.

Za ponazoritev vzemimo besedo *mula*, ki v slovenščini kot samostalnik moškega spola pomeni 'muslimanski verski učenjak in duhovni vodja', kot samostalnik ženskega spola pa 'domača žival' ter 'izražanje zamere, trme ali jeze' (*kuhati, pasti mulo*), redkeje 'dekle ali punca' (*mulci in mule*).[6] Če skušamo poiskati pomensko vez med vsemi omenjenimi pomeni oblike *mula* na podlagi podatkov v korpusu, lahko ugotovimo, da je samostojnost pomena 'muslimanski verski učenjak', npr.

- *ZDA menijo, da se vodja talibanov **mula** Omar in bin Laden skrivata v odročnih goratih predelih Pakistana ob meji z Afganistanom.*
- *Administracija predsednika Georgea Busha stopnjuje pritisk na iranski režim **mul**.*
- *"Prav ženske so tiste, ki so začele spodbijati absolutno avtoriteto iranskih **mul.**"*

določena tudi s kategorijo spola, s tem pa je zagotovljena tudi formalna prepoznavnost v osnovni obliki, ki se na ravni korpusa/leksikona pripisuje dvema lemama.[7] Na drugi strani je status pomenske povezanosti samostalnika *mula* za ženski spol med (a) 'žival', (b) 'čustven izraz' in (c) 'dekle' težje opredeljiv in odvisen od konteksta, posameznikove interpretacije in individualnih asociativnih povezav. Kaj od naštetega je mogoče razbrati že iz neposrednega konteksta, prikazujejo spodnji primeri, v katerih so prekrivni elementi sobesedila, kot jih prikazuje besedna skica v orodju Sketch Engine (Kilgarriff et al. 2004), za vse tri pomene podčrtani:

(a) 'žival'

- *Vsak, ki ima doma kakšno žival, vama bi vedel povedati, da se trmasta **mula** premakne v pravo smer le tedaj, ko ji ponujamo korenje.*
- *Poleg paše krav in **mul** sta podobno kot v preteklosti najbolj razširjeni ovčereja in kozjereja.*

(b) 'izražanje zamere, trme'

- *Namesto dialoga oboji pasejo **mulo**, ko čakajo na poteze nasprotne strani.*
- *Vsi smo že doživeli trenutke gledanja v tla ali jezno »kuhanje **mule** « in trmasto vztrajanje pri svojem.*

(c) 'dekle, punca'

- *Očitno so v zadnjem času moderni filmi, ki govorijo o upornih hčerkah, ki bi rade divjale, očki pa jim ne pustijo, vendar trmaste **mule** vseeno to dosežejo.*

6 V sodobni pisni slovenščini raba v tem pomenu ni pogosta, je pa vezana na sproščeno komunikacijo, zlasti med mladimi ali za ustvarjanje vtisa take komunikacije. Vezanost na zahodno narečje, kot predvideva SSKJ, v splošnem korpusu ni razvidna.

7 Analiza slovničnih relacij izpostavlja tudi nadpovprečno pogosto sopojavljanje z lastnimi imeni, npr. *Omar, Mohamed, Abdul* itd.

V takih situacijah, kot bomo prikazali na rešitvi konkretnega primera v SSKJ, se potrjuje ugotovitev, da princip ugotavljanja dobesednih in prenesenih (bodisi etimološko povezanih ali nepovezanih) pomenov ne zagotavlja enotnih (Taylor 2003), predvsem pa ne ustreznih rešitev z vidika odražanja jezikovne rabe v določeni jezikovni skupnosti.

3 HOMONIMIJA IN VEČPOMENSKOST V SLOVARJIH

Kot smo omenili v uvodu, je homonimija primarno slovarski problem, pri katerem gre v izhodišču za to, kako prikazati izrazno enoto v odnosu do njene vsebine v slovarju oz. kako povezati posamezne besedne oblike s posameznimi pomeni. Rešitve se nagibajo bodisi k (a) reševanju uporabniških problemov – v ta okvir sodijo zlasti slovarji za učenje tujega jezika in slovarji za šolsko populacijo ali določene uporabniške skupine – bodisi so (b) usmerjeni v čim bolj dosleden prenos teoretičnih izhodišč v slovarsko prakso, k čemur težijo zlasti akademijski in zgodovinski slovarji. Zato se upravičeno zastavlja vprašanje, ali prenos določene teorije v slovar vzdrži njegovo uporabniško naravnanost, v kolikšni meri je teoretična obravnava problema odvisna od logike slovarskega gesla, kot se je uveljavila zlasti v tiskanem mediju, in kateri princip zagotavlja zanesljivejšo slovarsko informacijo. Atkins in Rundell (2008: 192) navajata pet problematičnih situacij, ki jih znotraj homonimije obravnavajo slovarji:

(1) Identičen zapis in izgovor; različen pomen in etimologija, npr. nizi kot v SSKJ:

> **kóma**[1] -e ž med. *popolna, globoka nezavest:* bolnik je že dva dni v komi
> **kóma**[2] -e ž adm. *vejica, zlasti decimalna:* dve koma pet
> **kóma**[3] -e ž astr. *plinski ovoj, ki obdaja jedro kometa:* komet je imel okroglo komo in dolg rep

Gre za klasične primere homonimije, kot je v angleščini npr. *bank* v pomenu 'rečno obrežje' in v pomenu 'finančna ustanova'. Kot poudarjata Atkins in Rundell (2008: 192), se za ločeno obravnavo, tj. za navedbo homonimov, v takih primerih navadno odločajo zgodovinski in akademijski slovarji, vendar pa, kot še navajata, sodobna leksikografija tak princip opušča predvsem zaradi večje uporabniške naravnanosti in možnosti pojasnjevanja etimoloških podatkov na drugačen način, npr. v ločenih zavihkih (prim. Krek et al. 2013; Veliki slovar poljskega jezika[8]), saj se ugotavlja, da za uporabnike ugotavljanje etimološkega izhodišča ne more predstavljati temelja za organizacijo podatkov v slovarju in še manj izhodišča za dostop do iskane pomenske informacije (Moon 1987).

8 Wielki słownik języka polskiego: http://www.wsjp.pl/ (dostop 16. 7. 2015).

(2) Identičen zapis in izgovor; različen pomen. Ta kategorija zajema, če upoštevamo primere v SSKJ, dve podskupini. Prvo tvorijo (potencialni) homonimi, kjer prihaja pri osnovni obliki, npr. pri pridevnikih, do nanašanja na dve izhodiščno samostojni samostalniški lemi, kot prikazuje primer (a), drugo pa primeri, kjer se samostojnost lem ugotavlja na podlagi (sinhrone) pomenske neprekrivnosti, kot prikazuje primer (b):

(a) **bázičen**[1] -čna -o prid. *nanašajoč se na baza*[1]*:* bazična industrija
bázičen[2] -čna -o prid. *nanašajoč se na baza*[2]*:* bazična reakcija

(b) **víla**[1] -e ž *razkošneje grajena hiša z vrtom*
víla[2] -e ž mitol. *lepi, mladi ženski podobno bitje, ki živi v gozdu ali v vodi*

Slovarji, ki utemeljujejo homonimijo na ugotavljanju povsem ločenih pomenskih izhodišč, kamor sodi tudi SSKJ, obravnavajo take primere kot homonime, vendar pa, kljub nekaterim na prvi pogled očitno različnim pomenskim izhodiščem (npr. bazičen[1] in bazičen[2]), naletimo tudi na primere, kjer pomenske (ne)povezanosti ni mogoče jasno določiti.

Sivo polje med homonimijo in večpomenskostjo, kot prikazuje Tabela 1, potrjujejo tudi različne rešitve tovrstnih primerov v izbranih angleških enojezičnih slovarjih.

Za primerjavo smo vzeli samostalnik *punch*[9] v štirih enojezičnih angleških slovarjih, ki so prosto dostopni na spletu, pri čemer sta CDE in ODE namenjena primarno maternim govorcem, MED in CALD pa zlasti tujim govorcem. Kot je razvidno iz primerjave, se za prikaz v treh samostojnih iztočnicah (homonimih) odločata CED in ODE z nekoliko drugačno obravnavo pomena, ki se nanaša na 'ostrino, udarnost': v CED je obravnavan kot samostojni pomen, v ODE pa kot pod- oz. preneseni pomen. Uvrstitev drugih pomenov pod homonimne iztočnice je identična, le da ima CED kar štiri pomene, ki se navezujejo na orodje oz. napravo, ODE pa le dva. Pomen za 'vrsto alkoholne pijače' oba slovarja pripišeta samostojni homonimni iztočnici. Nasprotno obravnavata MED in CALD samostalnik *punch* kot eno samo iztočnico s štirimi samostojnimi pomeni, razlika je le v njihovi razvrstitvi.

9 Samostalnik ima v angleščini, ne glede na to, ali se obravnava kot homonim ali ne, navadno 4 pomene: 'udarec', 'ostrina, udarnost', 'alkoholna pijača' in 'orodje za luknjanje'.

Tabela 1: Različni pristopi k obravnavanju homonimije oz. večpomenskosti v izbranih angleških enojezičnih slovarjih.

Macmillan English Dictionary (MED)[10]	Collins English Dictionary (CED)[11]	Oxford Dictionary of English (ODE)[12]	Cambridge Advanced Learners Dictionary (CALD)[13]
večpomenskost	homonimija	homonimija	večpomenskost
punch (noun) **1** the action of hitting someone or something with your fist **2** a sweet drink made with fruit juice and usually alcohol **3** the emotional power of something such as a performance that affects how people feel **4** a tool for making a hole in something	**punch**[1] (noun)[14] **1** a blow with the fist **2** (informal) telling force, point, or vigour **punch**[2] **1** a tool or machine for piercing holes in a material **2** any of various tools used for knocking a bolt, rivet, etc, out of a hole **3** a tool or machine used for stamping a design on something or shaping it by impact **4** the solid die of a punching machine for cutting, stamping, or shaping material **5** *computing* a device, such as a card punch or tape punch, used for making holes in a card or paper tape **punch**[3] any mixed drink containing fruit juice and, usually, alcoholic liquor, generally hot and spiced	**punch**[1] (noun) **1** A blow with the fist **1.1** *informal* The strength needed to deliver a blow **1.2** *informal* The power to impress or attract attention; impact **punch**[2] **1** A device or machine for making holes in materials such as paper, leather, or metal **2** A tool or machine for impressing a design or stamping a die on a material **punch**[3] A drink made from wine or spirits mixed with water, fruit juices, spices, etc., and typically served hot	**punch** (noun) **1** a forceful hit with a fist **2** the power to be interesting and have a strong effect on people **3** a cold or hot drink made by mixing fruit juices, pieces of fruit, and often wine or other alcoholic drinks **4** piece of equipment that cuts holes in material by pushing piece of metal through it

10 http://www.macmillandictionary.com/ (dostop 16. 7. 2015).
11 http://www.collinsdictionary.com/dictionary/english (dostop 16. 7. 2015).
12 http://www.oxforddictionaries.com/ (dostop 16. 7. 2015).
13 http://dictionary.cambridge.org/ (dostop 16. 7. 2015).
14 Razlage za glagol in samostalnik so navedene znotraj posamezne homonimne iztočnice ločeno, najprej za glagol, nato za samostalnik.

Kot opozarjata Atkins in Rundell (2008: 192) je dejstvo, da leksikografi dejansko nimajo jasnih navodil, kako se odločati v posameznih primerih, poleg uporabniške naravnanosti slovarjev, tj. homonimija za materne govorce in večpomenskost za tuje, najpogosteje vzrok, da se nekateri slovarji odločajo za obravnavo znotraj ene iztočnice in ne dveh ali več samostojnih iztočnic.

(3) Identičen zapis; različen izgovor in pomen, npr. v SSKJ:

> **zástava** -e ž nar. primorsko *pas oblakov, megle nad gorami ob nastopu burje*
> **zastáva** -e ž 1. *kos tkanine določene barve ali več barv, ki predstavlja simbol kake države, naroda, organizacije* 2. ekspr., s prilastkom *nazor, program, gibanje:* zapustil je narodno zastavo
>
> **molíti** mólim nedov. 1. rel. *usmerjati misli, prošnje k Bogu* 2. *po božje častiti*
> **molíti** -ím nedov. 1. *imeti, ohranjati kaj v položaju, da seže bolj daleč kot sosednje stvari* 2. *hoteti dati, izročiti kaj, držeč pred seboj*

Atkins in Rundell (ibid.) ugotavljata, da je v večini sodobnih angleških slovarjev razlika v naglasnem mestu oz. sploh naglasni paradigmi avtomatično pogoj za obravnavo besede v samostojni iztočnici, tako da je v zvezi z vsako posebej mogoče v celoti predstaviti tudi oblikovno in naglasno paradigmo.

(4) Identičen zapis in izgovor; različen pomen in velika začetnica, npr. v Slovenskem pravopisu (SP):

> **átlas**[1] -a m *zbirka zemljevidov, slik iz določene stroke v obliki knjige:* izdati, sestaviti atlas
> **Átlas**[1] -a m, zem. i. |afriško gorovje|

Tudi v takih primerih se večina slovarjev odloča za navedbo samostojnih iztočnic, reševanje problema pa je v prvi vrsti odvisno od tega, ali slovar vključuje tudi prikaz lastnih imen.

(5) Identičen zapis in izgovor; različen pomen in kategorialne lastnosti. Atkins in Rundell (2008: 198) navajata v tem sklopu primere, ki so tipični predvsem za angleški jezik, v slovenščini pa je na tej podlagi izkazana razlika v obravnavi homonimije v SSKJ in drugih slovarskih priročnikih (gl. razdelka 3.2 in 3.3). Zlasti ko gre za besede z različno besednovrstno kategorijo, se slovarji navadno odločajo za obravnavo v samostojnih geslih oz. znotraj enega gesla ločeno (prim. Tabelo 1). V slovenščini sodijo v ta sklop primeri, kjer je identičnemu zapisu mogoče pripisati različno besedno vrsto, lahko pa tudi druge oblikoskladenjske lastnosti, kot so kategorija spola, sposobnost stopnjevanja pri pridevnikih itd. V SSKJ so taki primeri obravnavani kot samostojne iztočnice brez indeksne oznake, npr.

pŕst -a m 1. *vsak od petih gibljivih podaljškov dlani ali stopala* 2. *del rokavice, ki pokriva prst*

pŕst -í ž *vrhnja plast tal, ki vsebuje razkrojene organske snovi*

3.1 Homonimija vs. večpomenskost v SSKJ, SSKJ2 in NSSKJ

V zvezi z obravnavo homonimije v SSKJ lahko v Uvodu beremo: »Homonimna postavitev gesel se ne opira na etimologijo, temveč na stanje v sodobnem jeziku, s težnjo, da bi bilo homonimnih gesel v slovarju čim manj« (Uvod XVIII). Znotraj tega predvideva SSKJ dva tipa homonimnih iztočnic: iztočnice, ki se enako pišejo in pripadajo različnim besednim vrstam, in enako pisane iztočnice iste besedne vrste, ki so razvrščene po pogostosti in označene z nadpisano številko, npr.

tòp tôpa m, mn.	in	**kós**[1] -a m
tòp tôpa -o prid.		**kós**[2] -a m
tòp neskl. pril.		
tòp medm.		

Iz analize posameznih gesel je mogoče še ugotoviti, da so kot samostojne iztočnice obravnavane tudi enakopisnice, ki nimajo prekrivne naglasne paradigme, npr.

màr in már prisl.
mar prisl.

Obravnava homonimije je v SSKJ problematična predvsem v primerih, ko se skuša poiskati jasno določene meje pomenske (ne)povezanosti. Težnja po enotnih in konsistentnih teoretičnih rešitvah namreč lahko vodi v uniformno reševanje problema, ki ustvarja homonimna razmerja tudi v primeru razvidnih pomenskih asociacij, npr.

trôšnja[1] -e ž redko *trošenje, trosenje:* trošnja gnoja
trôšnja[2] -e ž zastar. *trošenje, poraba:* trošnja denarja

rotacíjski[1] -a -o prid. *nanašajoč se na rotacijo, vrtenje*
rotacíjski[2] -a -o prid. *nanašajoč se na rotacijo, stroj*

bliskavica[1] -e ž *oddaljeno bliskanje brez grmenja:* poletna bliskavica
bliskavica[2] -e ž fot. *priprava za močno trenutno osvetlitev pri fotografiranju:* fotoaparat ima vgrajeno bliskavico

Po drugi strani se problematika kaže v nekonsistentnih odločitvah pri obravnavi prenesenih pomenov, ki so ponekod predstavljeni kot večpomenskost (*pero,*[15] *izgovor*[16]), ponekod pa kot homonimi (*veda*):[17]

peró -ésa s
1. *kožna tvorba iz roževinastega tulca s pahljačastimi izrastki, ki v velikem številu pokriva telo ptic*
2. *majhna kovinska priprava s priostrenim koncem za pisanje, risanje*
3. publ., *s prilastkom pisatelj, književnik*
4. *vzmet*
5. star. *(rastlinski) list*

izgóvor -a s
1. *opravičilo, pri katerem se navadno ne navaja pravi, resnični vzrok*
2. *oblikovanje glasov, besed z govorilnimi organi*

véda[1] -e ž
1. *dejavnost, ki si prizadeva metodično priti do sistematično izpeljanih, urejenih in dokazljivih spoznanj*
2. zastar. *znanje*
3. zastar. *veščina, spretnost*

véda[2] -e ž nav. mn.
vsaka od štirih najstarejših knjig brahmanizma

Odločitve, ki jih leksikograf sprejema glede prepoznavne pomenske povezanosti v imenu celotne jezikovne skupnosti, se – tudi zaradi naslonitve na SSKJ – prenašajo tudi v Osnutek za izdelavo NSSKJ (Gliha Komac et al. 2015: 9), kjer je obravnava homonimije utemeljena takole:

> Za razliko od večpomenskosti, do katere prihaja zaradi jezikovnih potreb uporabnika /.../, gre pri homonimnih leksemih za popolno medsebojno vsebinsko nepovezanost, zato **homonimi ne morejo vzpostavljati enakovrstnih pomenskih povezav** (zlasti metonimičnih in metaforičnih), **kot jih pomeni večpomenskega leksema**. /.../ V NSSKJ zato homonime obravnavamo oz. prikazujemo v različnih slovarskih sestavkih, in to **kljub manjši avtomatiziranosti izdelave slovarskega sestavka** in posledično **večjemu deležu jezikoslovčeve dejavnosti** pri redakciji, saj se zdi to

15 Vidovič Muha (2000: 198) denimo v nasprotju s SSKJ pojmuje razmerje *pero* – 'del perja' in pero – 'pisalo' kot enakoizraznost (in ne kot večpomenskost), ki nastaja kot posledica jezikovnega razvoja.

16 V SSKJ je glede na SP in SSKJ razlika tudi v naglasu.

17 Kot je mogoče razbrati iz prostodostopnega etimološkega slovarja za angleščino http://www.etymonline.com/index.php?term=Veda&allowed_in_frame=0 (dostop 16. 7. 2015), se besedi stikata v staroindijskem korenu, ki pomeni 'vedenje, znanje; razumevanje'.

strokovno najkorektnejša odločitev, tudi zaradi kakovosti jeziko(slo)vnih podatkov, ki jih slovar ponuja **uporabniku**. (Poudarila P. G.)

V SSKJ2 je homonimija deloma drugače obravnavana kot v SSKJ, in sicer se briše razlikovanje med homonimi znotraj besednovrstnih kategorij in med različnimi besednovrstnimi kategorijami. Ta razlika je bila v SSKJ nakazana z nadpisano številko pri homonimih znotraj iste besednovrstne kategorije (gl. zgoraj primere *top* in *kos*).

3.2 Obravnava homonimije v slovarju sodobnega slovenskega jezika (SSSJ; primerjalno z obstoječimi spletnimi rešitvami)

Če si prikaz homonimije v odnosu do večpomenskosti ogledamo najprej z vidika načina predstavitve v obstoječih slovarjih, Osnutku za NSSKJ in v Predlogu za izdelavo SSSJ, lahko ugotovimo, da različne možnosti prikazovanja upoštevajo merila (a) izrazne prekrivnosti, (b) večpomenskosti oz. pomenske različnosti, (c) prepoznavanja kategorialnih oblikoskladenjskih lastnosti in (d) uporabo indeksiranja. Na tej podlagi zasnovane različne pristope lahko shematično prikažemo takole ('p'= pomen):

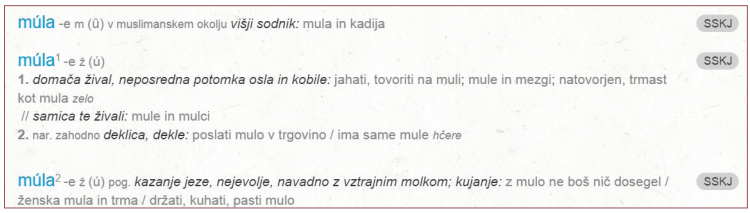
SSKJ	**SSKJ2/SP/NSSKJ**	**SSSJ**
mula sam. m. sp – 'p'	mula[1] sam. m. sp – 'p'	mula sam. m. sp – 'p'
mula[1] sam. ž. sp – 'p1'	mula[2] sam. ž. sp – 'p1'	mula sam. ž. sp – 'p1'
'p2'	'p2'	'p2'
mula[2] sam ž. sp – 'p'	mula[3] sam. ž. sp. – 'p'	'p3'

Razlike so očitne predvsem pri rešitvah SSSJ glede na druge slovarje, medtem ko se SSKJ od drugih inštitutskih slovarjev ločuje le v načinu indeksiranja homonimnih iztočnic.

3.2.1 SSKJ

Na slovarskem portalu ISJFR Fran, ki prenaša slovarske informacije knjižno zasnovanih slovarjev na splet,[18] je beseda *mula* prikazana v treh samostojnih iztočnicah, od katerih je prva za samostalnik moškega spola brez indeksne oznake, drugi dve, obe za samostalnik ženskega spola, pa imata indeksni oznaki 1 in 2 (Slika 1).

18 Vključeni so vsi slovarji razen SSKJ2, do katerega prek spleta ni mogoče prosto dostopati, hkrati pa ima tudi drugačno vizualizacijsko zasnovo.

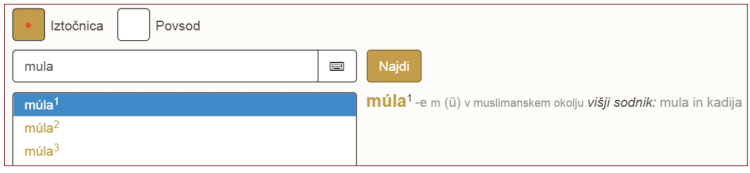


Slika 1: Prikaz homonimne iztočnice *mula* v SSKJ na slovarskem portalu Fran.

Prikaz zaporednih homonimnih iztočnic, kjer je neposredno ob iztočnici na voljo tudi pomenska razlaga, je z vidika dostopanja do informacij rešitev, ki je neproblematična v smislu, da uporabniku pri iskanju določenega pomena, ki ga v obliki izraza vnese v iskalno okence, ni treba poznati zgodovinskega razvoja besednega pomena, še preden ve, kaj beseda dejansko pomeni, saj je mogoče predvidevati, da informacije sicer ne bi iskal v slovarju (prim. Moon 1987: 88). Ker ima uporabnik ob zaporednih identičnih iztočnicah hkrati tudi dostop do pomenske vsebine, dejstvo, da so pomeni vezani na tri ločene enote, ni odločilno za to, da sploh pride do vsebine, ki ga zanima, saj je ta neposredno razvidna.

3.2.2 SSKJ2

V spletnem prikazu iste besede v SSKJ2 uporabnik dostopa do pomenskega opisa prek treh samostojnih iztočnic, ki se razlikujejo le v indeksni številki, pomen pa je prikazan samo pri prvi iztočnici, pri drugih dveh pa mora uporabnik preverjati vsebino z dodatnim poizvedovanjem. Tudi podatek o tem, da gre pri mula[1] za samostalnik moškega spola, pri mula[2] in mula[3] pa za samostalnik ženskega spola, na tej ravni ni razviden (Slika 2):



Slika 2: Prikaz homonimne iztočnice *mula* v SSKJ2 na spletu.

Tehnična rešitev, kot jo predvidevajo SSKJ, SSKJ2 in NSSKJ, ki je teoretično pogojena z ugotavljanjem pomenske – zgodovinske in sinhrone – (ne)prekrivnosti med pomenskimi sestavinami znotraj natančno strukturirane pomensko-sestavinske razlage, ima nujno tudi vsebinske posledice, ki se kažejo v pomenski informaciji, ki jo slovar daje uporabniku. Kot smo prikazali že na primerih gesel *pero, izgovor* in *veda* v SSKJ, je prikazana obravnava samostalnika *mula* z

indeksnima oznakama 2 in 3 problematična predvsem zato, ker sugerira pomensko povezavo med pomenoma 'domača žival' in 'deklica, dekle', ne pa denimo tudi med 'kazanje jeze' in 'domača žival', kar pa, kot smo prikazali v razdelku 2.2 z realnimi korpusnimi primeri, ne ustreza dejanskemu jezikovnemu stanju.

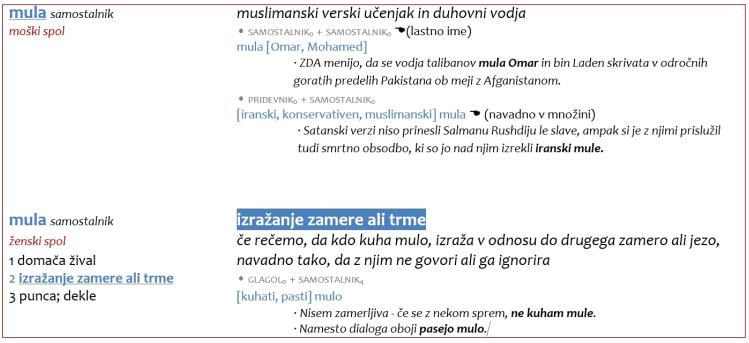
3.2.3 SSSJ

Homonimna razmerja, kot jih predvideva koncept SSSJ, za razliko od slovenske leksikalne teorije, kjer so vezana izključno na izrazno podobo jezikovnega znaka, potekajo med leksikalnimi enotami, tj. med besednimi pomeni in ne med besedami samimi (prim. Atkins in Rundell 2008: 162). Leksikalnih enot v SSSJ torej ne določa njihova izrazna podoba, ampak vsebina. Na ravni slovarske baze predstavlja iztočnica tehnični pojem, ki je usklajen s podatki v leksikonu oz. s postopkom lematizacije v korpusu, kjer je ločena obravnava iztočnic določena z oblikoslovnimi in skladenjskimi kategorijami, te pa so lahko pri prenosu v slovarsko bazo nadgrajene oz. spremenjene glede na slovarske potrebe. Z načinom obravnave homonimije v slovarski bazi pa so dana tudi izhodišča za vizualizacijo podatkov v spletnem slovarju, ki sledi potrebam uporabnikov in skuša prepoznavati njihove iskalne navade.

Ker postavljamo v SSSJ pomen v izhodišče slovarske obravnave in nanj pripenjamo tudi vse druge v slovarju pričakovane podatke, predstavljajo pomeni vozlišča v bazi strukturiranih podatkov na način, ki predvideva obravnavo v ločenih iztočnicah samo v primerih, kjer je različnost tudi formalno izpričana oz. kategorialno pogojena. Gre predvsem za primere, kjer prihaja do spremembe (a) kategorialnih lastnosti, kot so besedna vrsta, npr. *plesen* – sam. : *plesen* – prid., spol, npr. *prst* – sam. ž. spola : *prst* – sam. m. spola, in (b) v oblikoskladenjski, tj. spregatveni ali sklanjatveni, in naglasni paradigmi, npr. *pasti* – pasem : *pasti* – padem; *poročèn* – poročênega : *poróčen* – poróčnega. Med navedenimi oblikoslovno-glasoslovnimi spremembami je pri določanju homonimnih iztočnic mogoče upoštevati tudi razlike v oblikoslovni paradigmi, ki so povezane s spremembo pomenskih lastnosti, npr. z možnostjo stopnjevanja pridevnika zgolj v lastnostnem, npr. *bučen* – 'glasen': bolj bučen/bučnejši, ne pa tudi vrstnem pomenu: *bučen* – 'o buči', ob sicer eni sami pridevniški iztočnici.

Obravnava besede *mula*, kot je prikazana v nadaljevanju, je hipotetični prikaz samo ene od možnosti prikazovanja homonimije, ki temelji na zgoraj predstavljenih izhodiščih in je izdelana po vzoru Spletnega slovarja slovenskega jezika[19] (Slika 3):

19 Dostopno na: http://www.slovenscina.eu/spletni-slovar (dostop 16. 7. 2015).



Slika 3: Razmerje med homonimijo in večpomenskostjo v predlaganem SSSJ.

Kot je razvidno iz prikaza, uporabnik izbira med dvema sicer identičnima iztočnicama, pri čemer prva označuje samostalnik moškega, druga pa samostalnik ženskega spola. Ker je podatek razviden že na prvi ravni, lahko uporabnik loči med dvema kategorijama, ki vplivata na različno obnašanje besede v besedilu. Ker je samostalnik moškega spola v danem primeru enopomenski, je pomen razviden neposredno ob iztočnici, medtem ko je iztočnica, ki označuje samostalnik ženskega spola, večpomenska, osnovni namigi, ki uporabniku olajšajo izbiro najverjetnejšega pomena, pa so dani v pomenskem meniju v levem delu zaslona.

Prikaz obravnavanega primera v SSSJ temelji najprej na predpostavki, da lahko uporabnik na podlagi konteksta, v katerem je naletel na neznano besedo, s pomočjo pomenskih indikatorjev ter opozoril glede oblikovnih in naglasnih posebnosti že na ravni iztočnice ugotovi, kateri od navedenih pomenov je zanj zanimiv oz. v zvezi s katerim pomenom želi o besedi izvedeti več. Pri tem se v predlogu slovarja, ki je zasnovan za splet, odločamo za izpostavitev tistih lastnosti, zaradi katerih je obnašanje besede v iztočnici v besedilu različno in lahko predstavlja realen uporabniški problem. Taka lastnost je pri obravnavanem primeru kategorija spola, ki vpliva na variantnost besede v določenih oblikoskladenjskih položajih, npr. z možnostjo sklanjanja po moški ali ženski sklanjatvi: v rodilniku ednine: *talibanskega mula : talibanskega mule,* ali v imenovalniku množine: *iranski mule* : *iranske mule*.

V primeru samostalnika *mula* ženskega spola se, podobno kot v vseh primerih tega tipa, ne odločamo za iskanje povezav med osnovnimi pomeni, saj so pomenske asociacije uporabnikov različne, jasno določenih mej pa ni mogoče upravičiti niti teoretično. Zato pa toliko bolj težimo k pomenskemu opisu, ki čim bolj dosledno opisuje realno sinhrono rabo. Pri tem so tako leksikografom kot uporabnikom v veliko pomoč kolokacije, tipični besednozvezni in stavčni vzorci, pa tudi žanr, besedilni tip, področje rabe itd., ki odločilno vplivajo na naravno rabo besede v jeziku.

4 ZAKLJUČEK

Čeprav ločevanje homonimije (samostojnih iztočnic) od večpomenskosti (več pomenov ene iztočnice) ni realen sporočanjski, ampak slovarski problem, se učinkovitost reševanja skuša utemeljiti tudi teoretično, posledično pa naj bi bila bolj zanesljiva tudi slovarska informacija. Vendar pa se leksikografska interpretacija, ki uporablja etimologijo kot merilo razločevanja med homonimijo in večpomenskostjo, pogosto razlikuje od sinhrone govorčeve predstave, zato je tako izhodišče negotovo vodilo. Kot pravi Landau (2001), govorci namreč ne morejo nujno dojeti različnih pomenov kot povezanih med sabo, saj so etimološko različne besede včasih razvile podobne pomene, besede, ki jim je mogoče določiti isti izvor, pa so razvile različne pomene, tako da jih govorec sodobnega jezika ne dojema več kot sorodne.
Sodobna leksikografska praksa, zlasti slovarji, ki so izrazito uporabniško naravnani, so postopoma opustili homonimijo kot princip organiziranja iztočnic, s tem pa so sprožili tudi vprašanje, ali homonimija ostaja relevanten leksikografski koncept. Kot poudarjata Atkins in Rundell (2008: 281), je odgovor odvisen od tipa slovarja, njegovega namena in ciljnih uporabnikov. Mogoče je reči, da je navajanje izvora besed in njihovega razvoja osrednja funkcija zgodovinskih slovarjev, medtem ko je vloga homonimije z vidika sinhronega dojemanja pomenske vrednosti besed veliko bolj nejasna in nima ostro začrtanih prehodov. Tudi če uporabnik (npr. napačno vendar logično) predvideva, da je izražanje jeze ali trme, kot ga označuje zveza *pasti mulo,* povezano z značilnim obnašanjem te živali ali celo da obstaja podobnost med živaljo in izrazom na obrazu, bo najmanj zmeden, če ga bo slovar prepričeval, da take povezave v jeziku v resnici ni in hkrati navajal, da taka povezava nedvoumno obstaja med mulo kot živaljo in mulo kot dekletom. Strogo sledenje pravilom ločevanja homonimije od večpomenskosti pa povzroča tudi probleme pri učinkovitem dostopu do slovarske informacije. V primerih, ko slovar navaja obsežen homonimni niz (npr. v SSKJ2 ima beseda *tik* kar sedem zaporednih iztočnic), se moramo upravičeno vprašati, kako bo uporabnik našel ustrezni pomen ali stalno besedno zvezo, ki se skriva pod eno od navedenih iztočnic? Prav zaradi tega, vsaj kar se tiče sodobnih uporabniško usmerjenih slovarjev, homonimija ni več princip organizacije pomenov v odnosu do njihovih oblik v slovarju, ki bi bil uporabniku v pomoč.

V prispevku se zato zanašamo pri zapisovanju samostojnih iztočnic na formalne kriterije, ki so prek korpusa določeni v leksikonu besednih oblik oz. so prilagojeni za slovar v smislu, da upoštevajo kategorije, ki so zanimive za slovarski opis oz. so usmerjene v reševanje uporabniških problemov. Kategorije, ki se jih zdi smiselno pri tem upoštevati v slovarski bazi, so poleg besedne vrste še naglas in naglasno mesto, kategorija spola in živosti, sklanjatvena paradigma, možnost stopnjevanja

in variantnost. Pomeni, ki jih na podlagi teh kategorij »pokrivajo« iztočnice, so lahko v medsebojni pomenski povezavi, tako da so organizirani v podpomene, ali pa take povezave ni mogoče nedvoumno vzpostaviti, zato ostajajo osnovni pomeni relativno neodvisni drug od drugega.

V spletnem prikazu slovarskih podatkov je več izrazno prekrivnih iztočnic smiselno prikazovati hkrati s tistimi kategorijami, ki opozarjajo na razlike oz. posebnosti, ki nastanejo pri prenosu v besedilo. Večnivojskost prikaza podatkov, ki jo omogoča splet, je zato smiselno izkoristiti na način, ki bo prilagojen uporabniškemu obnašanju pri iskanju podatkov, na podlagi analiz njihovih potreb pa izpostaviti tiste slovarske opise, ki prinašajo odgovore na potencialna uporabniška vprašanja.

Leksikografska orodja za slovenščino: slovnica besednih skic

Simon Krek

Abstract
This paper describes the Word Sketch module in the Sketch Engine, a tool designed to help lexicographers interpret language data in large (grammatically annotated) text corpora. Word sketches are one-page automatic corpus-based summaries of a word's grammatical and collocational behaviour. The core of the module is sketch grammar, which uses regular expressions over part-of-speech tags to identify words in different syntactic relations in a given sentence. A word sketch represents a summary of all grammatical relations, and provides lists of statistically ranked collocations of a word in all given relations. This paper describes the sketch grammar for Slovenian as it was designed for the purposes of the compilation of Slovenian Lexical Database in the Communication in Slovenian project.

Keywords: lexicography, corpus analysis, collocations, grammatical relations, Sketch Engine

Ključne besede: leksikografija, korpusna analiza, kolokacije, slovnične relacije, Sketch Engine

1 UVOD

Namen prispevka je razmeroma tehničen opis delovanja modula Besedne skice *(Word sketches)*, ki je del leksikografskega orodja Sketch Engine. Orodje je bilo celostno že opisano v drugih prispevkih (Kilgarriff et al. 2004; Krek in Kilgarriff 2006; Yeroshina Pobirk et al. 2009; Gorjanc in Fišer 2013; Thomas 2015). Modul za poljubno korpusno lemo izdela besedno skico, avtomatsko generiran na korpusnih podatkih temelječ povzetek slovničnega in kolokacijskega vedenja neke besede. Besedna skica prikazuje leksikalni profil iskane leme s podatki o dvočlenskih in tročlenskih enotah, ki izkazujejo njeno tipično besedilno okolje. Prispevek opiše prilagoditev modula Besedne skice za slovenski jezik za potrebe izdelave Leksikalne baze za slovenščino (Gantar 2015) v okviru projekta Sporazumevanje v slovenskem jeziku.[1]

Temelj modula Besedne skice je **slovnica besednih skic**, sestavljena iz formaliziranih zapisov slovničnih relacij (angl. *grammatical relations*) ali **gramrelov**, ki določajo, kateri podatki (kolokacije in pripadajoči korpusni zgledi), ki jih mehanizem najde v korpusu, bodo upoštevani pri izpisu besedne skice. Priprava slovničnih relacij je jezikovnotehnološki postopek, za katerega je potrebno poznavanje tipičnih sobesedilnih pojavov v konkretnem jeziku ter načina jezikoslovnega označevanja korpusov, ki je uporabljen v posameznem korpusu. V nadaljevanju opišemo formalizem, ki ga uporablja modul za zapis slovničnih relacij, ter osnove sistema jezikoslovnega označevanja, ki je uporabljen v korpusu Gigafida (Logar et al. 2012).

2 SLOVNIČNE RELACIJE V SLOVNICI BESEDNIH SKIC

Slovnica Besednih skic (dalje SBS) uporablja formalni sistem poizvedovanja po vsebini atributov posamezne korpusne pojavnice s pomočjo regularnih izrazov. Uporabljeni formalizem je nekoliko razširjena verzija pogosto rabljenega jezika CQL (angl. *Corpus Query Language*), ki je bil za namen iskanja podatkov v jezikoslovno označenih korpusih izdelan na začetku 90-ih let (Christ 1994). V prispevku ne opisujemo temeljnih značilnosti jezika CQL oz. regularnih izrazov, npr. funkcije nadomestnih znakov in podobno, temveč se osredotočamo zgolj na tiste elemente razširjenega formalnega jezika, ki so specifični za modul Besedne skice oziroma orodje Sketch Engine.

Za lažje razumevanje delovanja mehanizma SBS najprej na kratko predstavimo način kodiranja informacij, ki jih uporablja modul v korpusu Gigafida (več o tem

1 Spletna povezava: http://www.slovenscina.eu/ (dostop 7. 8. 2015).

v Logar et al. 2012, 68–76; Grčar et al. 2012; Erjavec et al. 2015) in v orodju Sketch Engine. V procesu jezikoslovnega označevanja (tokenizacija, segmentacija, lematizacija in oblikoskladenjsko označevanje) so vse pojavnice v korpusu razdeljene na razred ločil, ki so kodirane v XML elementu <c> in razred besed, ki so kodirane v elementu <w>. Vsem besednim pojavnicam sta v procesu označevanja strojno pripisana dva metapodatka: lema in oblikoskladenjska oznaka, ki sta zakodirana kot atributa @lemma in @msd. Posamezna pojavnica v označenem korpusu Gigafida v jeziku XML TEI P6 (Erjavec et al. 2010) je izpisana na naslednji način:

```
<w lemma="packati" msd="Ggnn">packati</w>
```

Kot vsebina elementa <w> je izpisana besedna oblika, ki se pojavlja kot del izvornega korpusnega besedila, v atributu »lemma« je avtomatsko pripisana osnovna oblika – v konkretnem primeru oblika glagola v nedoločniku – ter v atributu »msd« oblikoskladenjska oznaka. Oznake so bile pripisane strojno s pomočjo označevalnika Obeliks (Grčar et al. 2012), ki uporablja nabor oznak JOS (Erjavec et al. 2010). V zgornjem primeru oznako »Ggnn« razvežemo na naslednji način:

Ggnn = (G) glagol; (g) vrsta=glavni; (n) vid=nedovršni; (n) oblika=nedoločnik.[2]

Orodje Sketch Engine za omenjene tri tipe informacij uporablja nekoliko drugačna poimenovanja. Atributi posamezne korpusne pojavnice so:

- **word**: besedna oblika oziroma zapis pojavnice, kakršen se nahaja v korpusu;
- **lemma**: osnovna oblika besede po specifikacijah JOS;
- **tag**: zapis oblikoskladenjske oznake po specifikacijah JOS.

V konkordančniku Sketch Engine so besedna oblika in oba atributa vidni na način, kot prikazuje Slika 1:

vsakič v obliki tankega, mastnega filma	packajo /Ggnstm/packati	tvoje lase. Namoči jo v blagi milnici,
poletje preživljajo doma, so včeraj neumorno	packali /Ggnd-mm/packati	in uživali Malčki v barvi do
razpoloženi pa so uporabili kar roke. Najraje	packam /Ggnspe/packati	s prstnimi barvami, je povedala Špela,

Slika 1: Prikaz atributov korpusne pojavnice v konkordančniku Sketch Engine.

Če bodisi pri poizvedovanju v konkordančniku ali pri pisanju slovnice besednih skic uporabljamo navedene tri tipe informacij, bomo torej v jeziku CQL torej uporabili atribute *word*, *lemma* in *tag*. Navajamo nekaj primerov poizvedb v jeziku CQL, s katerimi lahko iščemo po omenjenih atributih:

2 Nabor oznak je na spletni strani: http://nl.ijs.si/jos/msd/html-sl/index.html (dostop 7. 8. 2015).

- [word=«levega«] – poiščemo vse pojavnice z zaporedjem črk »levega«;
- [lemma=«most«] – poiščemo vse pojavnice z osnovno obliko samostalnika »most« (most, mosta, mostu ...);
- [tag=«Somei«] – poiščemo vse pojavnice, pri katerih so bile pri oblikoskladenjskem označevanju prepoznane naslednje lastnosti: *samostalnik, občni, moški spol, ednina, imenovalnik;*
- [tag=«S.*«] – poiščemo vse pojavnice, pri katerih je bila pri oblikoskladenjskem označevanju prepoznana lastnost *samostalnik;*
- [tag!=«S.*«] – iz poizvedbe izločimo vse pojavnice, pri katerih je bila pri oblikoskladenjskem označevanju prepoznana lastnost *samostalnik;*
- [] – označuje katerokoli pojavnico s kakršnokoli lastnostjo.

Poleg opisanega iskanja po vsebini atributov posamezne korpusne pojavnice z regularnimi izrazi razširjeni formalizem orodja Sketch Engine uporablja še štiri različne vrste **direktiv** za pet tipov gramatičnih relacij. Te so lahko:

- **enodelne**: iskani izraz vzpostavi razmerje z odvisnim izrazom, odnos je enosmeren; če poiščemo npr. samostalnik in s pomočjo SBS vzpostavimo razmerje s pridevnikom kot levim prilastkom, bo sistem pri samostalniških iskanjih izpisal niz pridevnikov, pri iskanju pridevnika pa samostalniški niz iz obratnega odnosa ne bo izpisan;
- **recipročne** (direktiva *DUAL*): iskani izraz vzpostavi razmerje z odvisnim izrazom, odnos je dvosmeren; če poiščemo npr. samostalnik in s pomočjo SBS vzpostavimo razmerje s pridevnikom kot levim prilastkom, bo sistem pri iskanju samostalnikov v relaciji izpisal niz pridevnikov, pri iskanju pridevnika pa samostalniški niz iz obratnega odnosa;
- **tridelne** (direktiva *TRINARY*): iskani izraz vzpostavi razmerje z odvisnim izrazom, vmes pa se pojavlja še tretji element, tipično predlog; če poiščemo npr. glagol in s pomočjo SBS vzpostavimo razmerje s samostalnikom v predložni zvezi kot odvisnim izrazom, bo sistem pri iskanjih glagolov v relaciji izpisal niz samostalnikov, ki se pojavljajo za določenim predlogom;
- **simetrične** (direktiva *SYMMETRIC*): iskani izraz vzpostavi razmerje z drugim izrazom, odnos je simetričen, enakovreden; če poiščemo npr. samostalnik in s pomočjo SBS vzpostavimo razmerje z drugim samostalnikom, ki se pojavlja za prirednim veznikom, bo sistem pri samostalniških iskanjih izpisal drug niz samostalnikov;
- **unarne** (direktiva *UNARY*): iskani izraz ne vzpostavlja razmerja do drugega elementa z izpisom niza, temveč le zabeleži nadpovprečno pojavljanje tega razmerja v korpusu.

V SBS je mogoče vključiti še nekatere dodatne direktive, ki pa jih v opisani slovnici besednih skic ne uporabljamo:[3] SEPARATEPAGE, COLLOC in CONSTRUCTION. Prva omogoča odpiranje besednih skic z direktivo TRINARY na novi (spletni) strani glede na tretji element, kar pomeni, da se bo npr. pri kombinaciji glagol+predlog+samostalnik vsaka kombinacija omenjenih elementov z različnim predlogom (*na*, *pod*, *v*, *pri* itd.) odprla na novi (spletni) strani. COLLOC uporablja nadomestne nize na pozicijah dveh drugih elementov, kar pomeni, da bodo pod posamezno gramatično relacijo, ki vsebuje to direktivo, izpisane kombinacije, ki ustrezajo izbranim nadomestnim nizom. Konkreten primer je denimo kombinacija predlog+samostalnik+predlog, ki pod tem gramrelom pri iskanem samostalniku »razmerje« statistično izpostavi kombinacijo v_do (»v razmerju do«) ali pri samostalniku »sklad« kombinacijo v_z (»v skladu z«). Z uporabo direktive CONSTRUCTION rezultat poizvedbe z uporabljenim gramrelom dobimo kot element v posebnem stolpcu znotraj besedne skice z imenom *Constructions*. Ta direktiva se torej uporablja za izpostavljanje statistično izstopajočih skladenjskih relacij pri iskanih lemah, ne pa za iskanje kolokacij, v katerih nastopa.

Posamezna slovnična relacija, ki generira en kolokacijski niz, vsebuje (vsaj) tri elemente:

- vrsto relacije (ena od direktiv, ki so naštete zgoraj oz. brez direktive);
- ime relacije, ki je poljubno;
- definicijo relacije, tj. zapis iskalnega pogoja v jeziku CQL.

S številkama 1, 2, ki jima sledi dvopičje, določamo, kateri del iskalnega pogoja definiramo kot »iskani element« (1:), ki ustreza temu, kar vpišemo v iskalno okence v modulu Besedne skice, in »kolokacijski element« (2:), ki generira kolokacijski niz. V primeru direktive DUAL lahko 1: in 2: prevzameta tudi obratni vlogi, v odvisnosti od tega, kaj vpišemo kot iskalni pogoj. Če je denimo 1: definiran kot pridevnik in 2: kot samostalnik, bo besedna skica proizvedla samostalniški kolokacijski niz, če bomo v iskalno okence vpisali pridevnik, bomo dobili obratni rezultat. Kot primer predstavljamo eno od preprostejših slovničnih relacij in njeno interpretacijo:

```
*DUAL
=v_rodil-s/s-koga-česa
      1:[tag="S.*"] 2:[tag="S...r.*"]
```

Slovnična relacija št. 21	
Opis delov slovnične relacije	**Vsebina slovnične relacije**
vrsta relacije =>	recipročna (*DUAL)
ime relacije =>	=v_rodil-s/s-koga-cesa
definicija relacije =>	1:[tag="S.*"] 2:[tag="S...r.*"]

3 Te direktive vsebuje SBS, ki je bila izdelana za avtomatsko luščenje podatkov iz korpusa (Krek 2012). https://trac.sketchengine.co.uk/attachment/wiki/SKEW-3/Program/Krek_SKEW- 3.pdf?format=raw (dostop 7. 8. 2015).

Interpretacija definicije relacije:

Deli definicije slovnične relacije	Interpretacija
1:[tag="S.*"]	prvi element relacije je samostalnik
2:[tag="S...r.*"]	drugi element relacije je samostalnik v rodilniku

Ker je relacija recipročna, bosta pri samostalniških iskanjih prikazana oba niza – prvi, kjer se ob iskanem samostalniku na desni strani pojavlja drug samostalnik v rodilniku, in drugi, kjer se pri iskanem samostalniku ob rodilniški obliki tega samostalnika pojavlja na levi strani drug samostalnik v katerikoli paradigmatski obliki. Rezultat, ki ga za lemo *ocenjevanje* prikaže ta slovnična relacija, je:

v_rodil-s	14,278	-8.1
vino	2,251	8.09
urejenost	240	7.94
med	242	7.32
znanje	1,614	7.14
salama	104	7.02
uspešnost	242	6.69
mesnatost	41	6.5
mošt	49	6.04
škoda	567	5.98
tveganje	265	5.97
primernost	45	5.93
kakovost	480	5.92
boniteta	41	5.73
žganje	46	5.56
cviček	38	5.44
ustreznost	38	5.26
dobrota	76	5.17
skladnost	35	5.11
invalidnost	27	5.09
delazmožnost	15	5.08
težavnost	21	5.06
kompetenca	24	4.98
pirh	20	4.96
učinkovitost	81	4.93
vzorec	142	4.68

s-koga-česa	6,157	-3.5
objektivnost	63	7.55
metodologija	119	7.24
didaktika	33	7.02
kriterij	280	6.58
metoda	227	5.41
merilo	185	5.4
šampion	30	5.18
moderacija	7	5.14
pravičnost	35	5.14
rezultat	398	4.64
zanesljivost	17	4.62
strogost	7	4.56
način	603	4.35
proces	182	4.18
standard	88	4.06
postopek	376	3.9
publikacija	18	3.86
pravilo	117	3.81
akcija	159	3.75
pričetek	7	3.52
potek	30	3.47
sistem	392	3.47
organizator	42	3.35
pravilnost	7	3.31
legenda	19	3.25

Slika 2: Vzorčna slovnična relacija v Besedni skici.

Kombinacijo pri razmerju **'v_rodil-s'** moramo torej brati na naslednji način:

ocenjevanje + **v_rodil-s** (*znanja, urejenosti, kakovosti,...*)
ocenjevanje znanja
ocenjevanje urejenosti
ocenjevanje kakovosti
...

Kombinacijo pri razmerju **'s-koga-cesa'** moramo brati:

s = kriterij, način, metodologija,... + **koga-česa** = *ocenjevanja*
kriterij ocenjevanja
način ocenjevanja
metodologija ocenjevanja
...

Primeri konkordanc pri tej slovnični relaciji:

v_rodil-s:
1. element: lema *ocenjevanje* v kateremkoli sklonu,
2. element: samostalniški kolokator v rodilniku:

kolokator: **kakovost**

»Skupna ocena po evropskem modelu	ocenjevanja	*kakovosti* v javnem sektorju v naši upravni
slovesnosti izročila direktorica dejavnosti	ocenjevanja	*kakovosti* pri Slovenskem inštitutu za kakovost
nameravamo spremeniti pristopa pri primerjalnem	ocenjevanju	*kakovosti* izdelkov in storitev. Znak VIP
ocenjen. Cena izdelka ni eden od kriterijev za	ocenjevanje	*kakovosti* . Izdelke bodo lahko z

kolokator: **škoda**

upravnico, ta ni imela pojma o kakršnemkoli	ocenjevanju	*škode* , povedala mu je le, da jo je pred
ekonomičnost (predimenzioniranje) konstrukcije. Med	ocenjevanjem	*škode* na porušenih mostovih po 2. svetovni
težiščna aktivnost zbiranje podatkov in	ocenjevanje	*škode* po suši ter izdelava analize oskrbe
vodenje sil za zaščito in reševanje, pomoč in	ocenjevanje	*škode* ob naravnih nesrečah ter organizacijo

s-koga-česa:
1. element: kolokator v kateremkoli sklonu,
2. element: lema *ocenjevanje* v rodilniku:

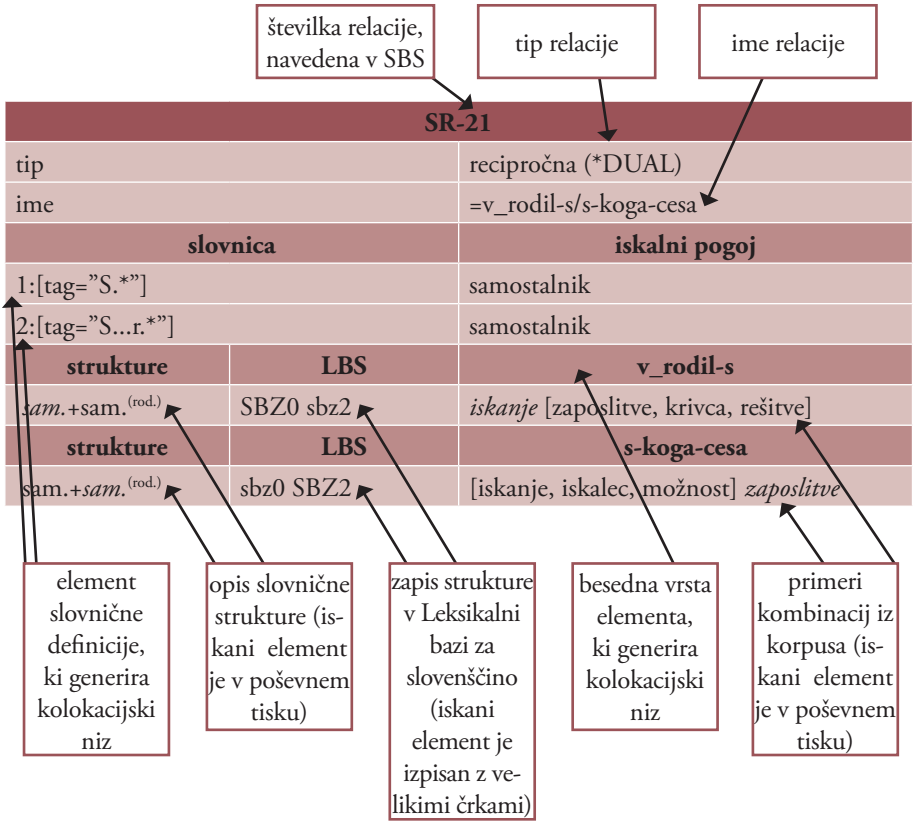
kolokator: **kriterij**

ocenjevanja oz. preverjanja *kriterijev*	ocenjevanja	trofej, ki sem jih takrat predlagal. Brez
športom glede popularnosti, pa tudi *kriteriji*	ocenjevanja	so kar pravi. In tisti, ki smo doživeli
lahko pokaže učni načrt, razloži *kriterij*	ocenjevanja	in podobno. A del demokracije je tudi prevzemanje
komisija je povedala temeljne *kriterije*	ocenjevanja	razpisa, ki pa jih je bilo treba čitati

kolokator: **metoda**

seznam trofej in idealno izdelano	***metodo***	**ocenjevanja** , če bi že v letu 1993 napravili prve korake
slovaško in švicarsko vino. Nova	***metoda***	**ocenjevanja** - pozitivna, pri tej je maksimalno število
vrednote igrajo pomembno vlogo. Izbira	***metode***	**ocenjevanja** , na primer odločitev med OLS in najmanjšo
preverjanja), pač pa sta tako vsebina kot	***metoda***	**ocenjevanja** (npr. način pridobivanja ustne ocene ustno
vsebin iz maturitetnega kataloga in	***metod***	**ocenjevanja** . S takšno oceno je kandidat seznanjen pred

V nadaljevanju na podoben način razlagamo 32 slovničnih relacij v opisani SBS, z naslednjimi podatki:



Slika 3: podatki o slovničnih relacijah v SBS.

Kratice, ki jih uporabljamo pri opisu slovničnih relacij:

kratica	razlaga	kratica	razlaga
im.	imenovalnik	**sam.**	samostalnik
rod.	rodilnik	**prid.**	pridevnik
daj.	dajalnik	**glag.**	glagol
tož.	tožilnik	**prisl.**	prislov
mest.	mestnik	**predl.**	predlog
orod.	orodnik	**vezn. elem.**	vezniški element

SR-01		
tip		recipročna (*DUAL)
ime		=kaksen?/kdo-kaj?
slovnica		**iskalni pogoj**
1: [tag="S.*"]		samostalnik
2: [tag="P.*"]		pridevnik
strukture	**LBS**	**kaksen?**
prid.+*sam.*	pbz0 SBZ0	[boleč, lep] *spomin*
strukture	**LBS**	**kdo-kaj?**
prid.+sam.	PBZ0 sbz0	*rdeča* [žoga]

SR-02		
tip		recipročna (*DUAL)
ime		=kako-kdaj?/je_pred
slovnica		**iskalni pogoj**
1:[tag="[PG].*"]		pridevnik, glagol
2:[tag="R.*"]		prislov
strukture	**LBS**	**kako-kdaj?**
prisl.+*glag.*	rbz GBZ	[nesrečno] *pasti*, [neznosno] *boleti*
prisl.+*prid.*	rbz PBZ	[kristalno] *čist*; [neznansko] *vesel*
strukture	**LBS**	**je_pred**
prisl.+glag.	RBZ gbz	*močno* [deževati], *drastično* [se znižati]
prisl.+prid.	RBZ pbz	*bledo* [rdeč], *hudo* [zapit]

SR-03		
tip		enodelna
ime		=veznik
slovnica		**iskalni pogoj**
1:[tag="[GPRS].*"]		glagol, pridevnik, prislov, samostalnik
strukture	**LBS**	**veznik**
prid.+vejica+vezn. elem.	PBZ0 Odv-da	*vesel*, da je kaj
sam.+vejica+vezn. elem.	SBZ0 Odv-ali	*vprašanje*, ali …
prisl.+vejica+vezn. elem.	RBZ Odv-da	*preprosto*, da (bolj ne more biti)
glag.+vejica+vezn. elem.	GBZ Odv-da	*vedeti*, da ...

SR-04		
tip		enodelna
ime		=predlog
slovnica		**iskalni pogoj**
1:[]		glagol, pridevnik, prislov, samostalnik
strukture	**LBS**	**predlog**
predl.+*prid.*	-	kot *nor*
predl.+*prisl.*	-	na *bolje*
predl.+*sam.*	-	pred *časom*

SR-05		
tip		enodelna
ime		=predl-za
slovnica		**iskalni pogoj**
1:[]		glagol, pridevnik, prislov, samostalnik
strukture	**LBS**	**predl-za**
glag.+predl.	-	*verjeti* v
prid.+predl.	-	*zanimiv* za
prisl.+predl.	-	*najpozneje* do
sam.+predl.	-	*voda* iz

SR-06		
tip		tridelna (*TRINARY)
ime		=%s
slovnica		**iskalni pogoj**
1:[tag="G.*"]		glagol
strukture	**LBS**	**=%s**
glag.+predl.+sam.(rod.)	vezljivostni vzorec[4]	*prilesti* do [polfinala]
glag.+predl.+sam.(daj.)	vezljivostni vzorec	*leteti* proti [vratnici]
glag.+predl.+sam.(tož.)	vezljivostni vzorec	*meriti* na [kolajno, stopničke]
glag.+predl.+sam.(mest.)	vezljivostni vzorec	*pogajati se* o [vdaji, izpustitvi]
glag.+predl.+sam.(orod.)	vezljivostni vzorec	*stisniti* med [prsti]

SR-07		
tip		tridelna (*TRINARY)
ime		=%s
slovnica		**iskalni pogoj**
1:[tag="S.*"]		samostalnik
strukture	**LBS**	**=%s**
sam.+predl.+sam.(rod.)	SBZ0 okoli sbz2	*ograja* okoli [hiše]
sam.+predl.+sam.(daj.)	SBZ0 proti sbz3	*ukrep* proti [kršiteljem]
sam.+predl.+sam.(tož.)	SBZ0 na sbz4	*spomin* na [otroštvo, mladost]
sam.+predl.+sam.(mest.)	SBZ0 ob sbz5	*incident* ob [meji]
sam.+predl.+sam.(orod.)	SBZ0 pred sbz6	*strah* pred [neuspehom]

SR-08		
tip		tridelna (*TRINARY)
ime		=%s_X
slovnica		**iskalni pogoj**
1:[tag="S.*"]		samostalnik
strukture	**LBS**	**%s_X**
sam.+predl.+*sam.*(rod.)	sbz0 do SBZ2	[čas] do *volitev*
sam.+predl.+*sam.*(daj.)	sbz0 k SBZ23	[molitev] k *bogu*
sam.+predl.+*sam.*(tož.)	sbz0 čez SBZ4	[most] čez *reko*
sam.+predl.+*sam.*(mest.)	sbz0 čez SBZ5	[padec] v *brezno*
sam.+predl.+*sam.*(orod.)	sbz0 nad SBZ6	[skrbništvo] nad *otrokom*

4 Pri glagolskih iztočnicah so pri kombinacijah s samostalniškimi in predložnimi zvezami v Leksikalni bazi zabeleženi vezljivostni vzorci, ne pa skladenjske strukture. Več o tem v Gantar (2015).

SR-09		
tip		tridelna (*TRINARY)
ime		=%s_X
slovnica		**iskalni pogoj**
1:[tag=”S.*”]		samostalnik
strukture	**LBS**	**%s_X**
glag.+predl.+*sam.*(rod.)	gbz preko SBZ2	[stopiti] preko *potoka*
glag.+predl.+*sam.*(daj.)	gbz k SBZ3	[privezati] k *postelji*
glag.+predl.+*sam.*(tož.)	gbz na SBZ4	[pomisliti] na *otroštvo*
glag.+predl.+*sam.*(mest.)	gbz pri SBZ5	[ostati] pri *življenju*
glag.+predl.+*sam.*(orod.)	gbz s SBZ6	[poriniti] s *silo*

SR-10		
tip		recipročna (*DUAL)
ime		=koga-cesa/v_rodil
slovnica		**iskalni pogoj**
1:[tag=”S...r.*”]		samostalnik
2:[tag=”Gg.*”]		glagol
strukture	**LBS**	**koga-cesa**
glag.+sam.(rod.)	vezljivostni vzorec	*izogibati se* [stikov] *bati se* [maščevanja]
strukture	**LBS**	**v_rodil**
glag.+*sam.*(rod.)	gbz SBZ2	[manjkati] *denarja* [braniti se] *očitkov*

SR-11		
tip		recipročna (*DUAL)
ime		=komu-cemu/v_dajal
slovnica		**iskalni pogoj**
1:[tag=”So..d.*”]		samostalnik
2:[tag=”Gg.*”]		glagol
strukture	**LBS**	**komu-cemu**
glag.+sam.(daj.)	vezljivostni vzorec	*izogibati se* [naporu] *uiti* [smrti]
strukture	**LBS**	**v_dajal**
glag.+*sam.*(daj.)	gbz SBZ3	[predajati se] *spominom* [odpovedati se] *ljubezni*

SR-12		
tip		recipročna (*DUAL)
ime		=koga-kaj/v_tozil
slovnica		**iskalni pogoj**
1:[tag="So..t.*"]		samostalnik
2:[tag="Gg.*"]		glagol
strukture	**LBS**	**koga-kaj**
glag.+sam.$^{(tož.)}$	vezljivostni vzorec	*videti* [priložnost] *sezuti* [škornje]
strukture	**LBS**	**koga-kaj**
glag.+*sam.*$^{(tož.)}$	gbz SBZ4	[zapravljati, izgubljati] *čas* [izgubiti] *nedolžnost*

SR-13		
tip		recipročna (*DUAL)
ime		=osebek_od/osebek_je
slovnica		**iskalni pogoj**
1:[tag="So..i.*"]		samostalnik
2:[tag=" Gg.st.*"]		glagol
strukture	**LBS**	**osebek_od**
sam.$^{(im.)}$+glag.	SBZ1 gbz	*spomini* [oživijo] *zastava* [visi]
strukture	**LBS**	**osebek_je**
sam.$^{(im.)}$+*glag.*	vezljivostni vzorec	[sodišče] *verjame* [odbor] *se sestane*

SR-14		
tip		enodelna
ime		=nedolocnik
slovnica		**iskalni pogoj**
1:[]		glagol, pridevnik, prislov, samostalnik
strukture	**LBS**	**nedolocnik**
glag.+glag.$^{(nedol.)}$	GBZ Inf-gbz	*uspeti* [doseči]
sam.+glag.$^{(nedol.)}$	SBZ0 Inf-gbz	(imeti) *možnost* [pritožiti se]
prid.+glag.$^{(nedol.)}$	PBZ0 Inf-gbz	*pripravljen* [oditi]; *sposoben* [izpeljati]
prisl.+glag.$^{(nedol.)}$	RBZ Inf-gbz	*dobro* [izkoristiti]

SR-15		
tip		enodelna
ime		=GPRS-inf
slovnica		**iskalni pogoj**
1:[tag=”G..n.*”]		glagol
strukture	**LBS**	**GPRS-inf**
glag.+*glag.*(nedol.)	gbz Inf-GBZ	[znati] *pasti*, [nehati] *boleti*, [morati] *paziti*
prisl.+*glag.*(nedol.)	rbz Inf-GBZ	[treba] *vedeti* [pomembno, potrebno] *poznati*
prid.+*glag.*(nedol.)	pbz0 Inf-GBZ	[dolžen] *upoštevati*
sam.+*glag.*(nedol.)	sbz1 Inf-GBZ	(biti) [užitek, čast] *igrati*

SR-16		
tip		recipročna (*DUAL)
ime		=zanikan/z_nikalnim
slovnica		**iskalni pogoj**
1:[tag=”So..r”]		samostalnik
2:[tag=”Gg.*”]		glagol
strukture	**LBS**	**zanikan**
glag.(neg.)+*sam.*(rod.)	Neg-gbz SBZ2	[ne izgubljati, ne zapravljati] *časa*
strukture	**LBS**	**z_nikalnim**
glag.(neg.)+sam.(rod.)	Neg-GBZ sbz2	*ne najti* [dokazov]

SR-17		
tip		recipročna (*DUAL)
ime		=kolicina_ob-s/kolicinski
slovnica		**iskalni pogoj**
1:[tag=”R.*”]		prislov
2:[tag=”S...r.*”]		samostalnik
strukture	**LBS**	**kolicinski**
prisl.+*sam.*(rod.)	rbz SBZ2	[nekaj, precej] *časa*
strukture	**LBS**	**kolicina_ob-s**
prisl.+sam.(rod.)	RBZ sbz2	*precej* [pozornosti, denarja]

SR-18		
tip		recipročna (*DUAL)
ime		=s_prislovom/s_prislovom
slovnica		**iskalni pogoj**
1:[tag="So..r.*"]		samostalnik
2:[tag="Gg.*"]		glagol
strukture	**LBS**	**s_prislovom**
glag.+prisl.+*sam.*(rod.)	gbz rbz SBZ2	[porabiti, preživeti] (veliko, precej) *časa*
strukture	**LBS**	**s_prislovom**
glag.+prisl.+sam.(rod.)	GBZ rbz sbz2	*kazati* (veliko, precej) [zanimanja]

SR-19		
tip		recipročna (*DUAL)
ime		=kaksen-p?/kaksen-g?
slovnica		**iskalni pogoj**
1:[tag="Gg.*"]		glagol
2:[tag="P....i.*"]		pridevnik
strukture	**LBS**	**kaksen-p?**
prid.(im.)+*glag.*	pbz1 GBZ	[vesel] *razlagati*
glag.+prid.(im.)	GBZ pbz1	*izgledati* [utrujen] *izpasti* [neumen]
strukture	**LBS**	**kaksen-g?**
glag.+*prid.*(im.)	gbz PBZ1	[vstati] *neprespan* [ostati, postati, zdeti se] *vesel*

SR-20		
tip		recipročna (*DUAL)
ime		=kaksnega-p/kaksnega-g?
slovnica		**iskalni pogoj**
1:[tag="Gg.*"]		glagol
2:[tag="P....r.*"]		pridevnik
strukture	**LBS**	**kaksnega-p**
glag.+prid.(rod.)	GBZ pbz2	*čutiti se* [osamljenega]
strukture	**LBS**	**kaksnega-g?**
glag.+*prid.*(rod.)	gbz PBZ2	[čutiti se] *veselega*

SR-21		
tip		recipročna (*DUAL)
ime		=v_rodil-s/s-koga-cesa
slovnica		**iskalni pogoj**
1:[tag="S.*"]		samostalnik
2:[tag="S...r.*"]		samostalnik
strukture	**LBS**	**v_rodil-s**
sam.+sam.(rod.)	SBZ0 sbz2	*iskanje* [zaposlitve, krivca, rešitve]
strukture	**LBS**	**s-koga-cesa**
sam.+*sam.*(rod.)	sbz0 SBZ2	[iskanje, iskalec, možnost] *zaposlitve*

SR-22		
tip		recipročna (*DUAL)
ime		=/oba-v-rod
slovnica		**iskalni pogoj**
1:[tag="P....r.*"]		pridevnik
2:[tag="S...r.*"]		samostalnik
strukture	**LBS**	**oba-v-rod**
prid.(rod.)+*sam.*(rod.)	pbz2 SBZ2	[šibkega] *zdravja*
strukture	**LBS**	**oba-v-rod**
prid(rod.)+sam.(rod.)	PBZ2 sbz2	*bledega* [obraza]

SR-23		
tip		tridelna (*TRINARY)
ime		=%s
slovnica		**iskalni pogoj**
1:[tag="P.*"]		pridevnik
strukture	**LBS**	**%s**
prid.+predl.+sam.(rod.)	PBZ0 glede sbz2 PBZ0 od sbz2	*negotov* glede [meril] *enoten* glede [sodelovanja] *zaripel* od [napora] *bled* od [jeze] *odrezan* od [sveta]
prid.+predl.+sam.(daj.)	PBZ0 k sbz3	*nagnjen* k [pretiravanju, debelosti]
prid.+predl.+sam.(tož.)	PBZ0 na sbz4	*pogolten* na [denar] *prijeten* na [pogled]
prid.+predl.+sam.(mest.)	PBZ0 po sbz5 PBZ0 pri sbz5	*velik* po [obsegu] *zasačen* pri [kraji]
prid. +predl.+sam.(orod.)	PBZ0 z sbz6	*zadovoljen* z [izkupičkom, razpletom]

SR-24		
tip		recipročna (*DUAL)
ime		=v_rodil-p/p-koga-cesa
slovnica		**iskalni pogoj**
1:[tag="P....i.*"]		pridevnik
2:[tag="S...r.*"]		samostalnik
strukture	**LBS**	**v_rodil-p**
prid.+sam.(rod.)	PBZ0 sbz2	*vreden* [spoštovanja, omembe]
strukture	**LBS**	**p-koga-cesa**
prid.+*sam.*(rod.)	pbz0 SBZ2	[poln, vreden] *denarja*

SR-25		
tip		recipročna (*DUAL)
ime		=v_dajal-p/p-komu-cemu
slovnica		**iskalni pogoj**
1:[tag="P....i.*"]		pridevnik
2:[tag="S...d.*"]		samostalnik
strukture	**LBS**	**v_dajal-p**
prid.+sam.(daj.)	PBZ0 sbz3	*zvest* [prijatelju, tradiciji]
strukture	**LBS**	**p-komu-cemu**
prid.+*sam.*(daj.)	pbz0 SBZ3	[nevaren] *državi*

SR-26		
tip		recipročna (*DUAL)
ime		=v_tozil-p/p-koga-kaj
slovnica		**iskalni pogoj**
1:[tag="P....i.*"]		pridevnik
2:[tag="S...t.*"]		samostalnik
strukture	**LBS**	**v_tozil-p**
prid.+sam.(tož.)	PBZ0 sbz4	*vreden* [milijardo]
strukture	**LBS**	**p-koga-kaj**
prid.+*sam.*(tož.)	pbz0 SBZ4	[dolžen] *vsoto*

SR-27		
tip		recipročna (*DUAL)
ime		=biti_kaksen?/osebek+biti
slovnica		**iskalni pogoj**
1:[tag="S...i.*"]		samostalnik
2:[tag="P....i.*"]		pridevnik
strukture	**LBS**	**biti_kaksen?**
sam.(im.)+biti+prid.(im.)	SBZ1 Vez-gbz pbz1	*spomin* je [boleč, živ]
strukture	**LBS**	**osebek+biti**
sam.(im.)+biti+*prid.*(im.)	sbz1 Vez-gbz PBZ1	[situacija, postrežba] je *slaba*

SR-28		
tip		simetrična (*SYMMETRIC)
ime		=
slovnica		**iskalni pogoj**
1:[tag="[GPRS].*"]		glagol, pridevnik, prislov, samostalnik
2:[tag="[GPRS].*"]		glagol, pridevnik, prislov, samostalnik
strukture	**LBS**	**priredje**
glag.+vez.+glag.	GBZ in gbz gbz in GBZ GBZ ali gbz	*pasti* in se poškodovati *spotakniti se* in pasti *kupiti* ali najeti
sam.+vez.+sam.	SBZ0 in sbz0	*čas* in denar
prid.+vez.+prid.	PBZ0 in pbz0	*hladen* in vlažen
prisl.+vez.+prisl.	RBZ in rbz	*sončno* in vroče

SR-29		
tip		recipročna (*DUAL)
ime		=s_trpnikom/trpnik
slovnica		**iskalni pogoj**
1:[tag="Pd.*"]		pridevnik
2:[tag="S...i.*"]		samostalnik
strukture	**LBS**	**s_trpnikom**
sam.(im.)+biti+*prid.* *(dežž.)*	sbz1 Vez-gbz PBZ1	[stavba, hiša, cerkev] je *zgrajena*
strukture	**LBS**	**trpnik**
sam.(im.)+biti+prid. (dežž.)	SBZ1 Vez-gbz pbz1	*cesta* je [speljana, zaprta, asfaltirana]

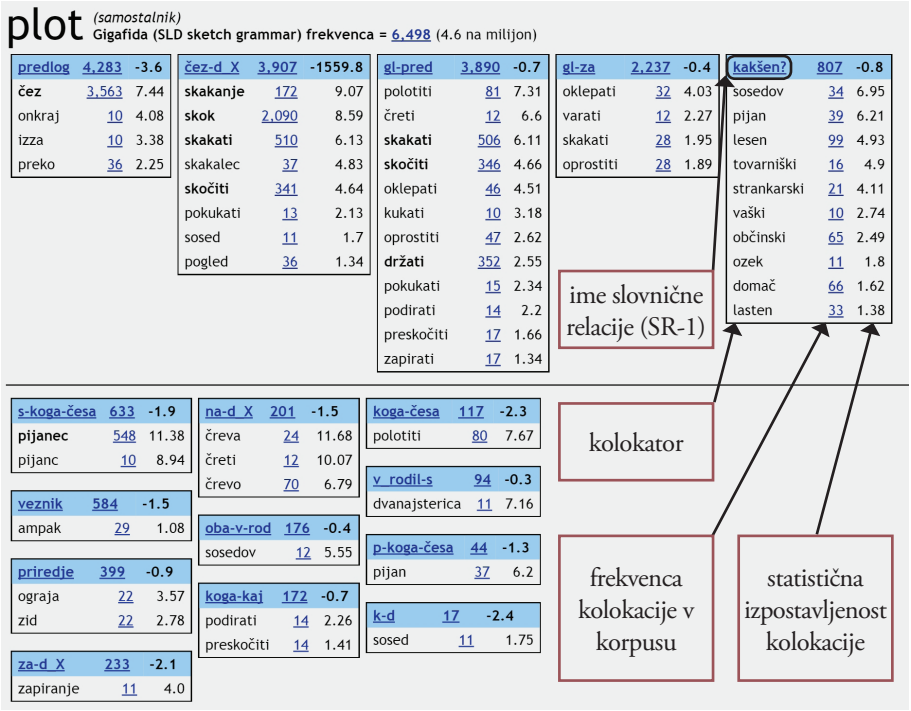
SR-30		
tip		enodelna
ime		=gl-pred
slovnica		**iskalni pogoj**
1:[]		glagol, pridevnik, prislov, samostalnik
strukture	**LBS**	**gl-pred**
glag.+*glag.*	-	[hoteti] *videti*
glag.+*prid.*	-	[jesti] *lačen*
glag.+*prisl.*	-	[igrati] *slabo*
glag.+*sam.*	-	[poučevati] *violino*

SR-31		
tip		enodelna
ime		=gl-za
slovnica		**iskalni pogoj**
1:[]		glagol, pridevnik, prislov, samostalnik
strukture	**LBS**	**gl-za**
glag.+glag.	-	*videti* [slišati, občutiti]
prid.+glag.	-	*veljaven* [predpisovati, obdavčiti]
prisl.+glag.	-	*vidno* [označiti, razočarati, pretresti]
sam.+glag.	-	*časopis* [objaviti, objavljati, poročati]

SR-32		
tip		unarna (*UNARY)
ime		=stev-pred
slovnica		**iskalni pogoj**
1:[tag="S.*"]		samostalnik
strukture	**LBS**	**stev-pred**
-	-	mesto, dan, mesec

Od navedenih 32 slovničnih relacij ali gramrelov jih 27 proizvede rezultat za samostalniške poizvedbe, 18 za glagolske, 17 za pridevnike in 9 za prislovne poizvedbe.[5] Na sliki 3 je kot primer prikazan rezultat samostalniške poizvedbe za lemo »plot«:

5 Samostalniške so vse razen SR-02, SR-06, SR-15, SR-19, SR-20 in SR-23, glagolske: SR-02, SR-03, SR-05, SR-06, SR-10, SR-11, SR-12, SR-13, SR-14, SR-15, SR-16, SR-18, SR-19, SR-20, SR-28, SR-29, SR-30, SR-31, pridevniške: SR-01, SR-02, SR-03, SR-04, SR-05, SR-14, SR-19, SR-20, SR-22, SR-23, SR-24, SR-25, SR-26, SR-27, SR-28, SR-30, SR-31, prislovne: SR-02, SR-03, SR-04, SR-05, SR-14, SR-17, SR-28, SR-30, SR-31.



Kot smo omenili, je bila predstavljena slovnica besednih skic izdelana za uporabo pri gradnji Leksikalne baze za slovenščino (LBS), ki je podrobno predstavljena v Gantar (2015; razmerje med SBS in LBS je opisano predvsem v poglavju »Skladenjske strukture«). Relacije, ki po presoji leksikografa proizvajajo relevantne kolokacijske nize, so bile v leksikalni bazi beležene pod predvidenimi skladenjskimi strukturami s poenotenim zapisom, kakršen je naveden zgoraj pri opisu posameznih elementov SBS (npr. SBZ2 = samostalniška besedna zveza s samostalniškim jedrom v rodilniku). Ta verzija je bila torej pripravljena za specifičen namen beleženja čim večjega števila skladenjskih odnosov s podrejenimi kolokacijskimi podatki in korpusnimi zgledi v okviru predvidenega opisa slovenščine v LBS. Širše gledano je mehanizem SBS smiselno dojemati projektno. Slovnice je mogoče prilagoditi potrebam različnih projektov, kar v resnici kaže tudi dosedanja zgodovina omenjenih slovnic za slovenščino.

Opisana verzija je druga pomembnejša verzija SBS za slovenščino. Prva (Krek in Kilgarriff 2006) je bila izdelana poskusno, njen namen bil predvsem preizkusiti mehanizem, hkrati pa ohranja mednarodni značaj, saj so imena slovničnih relacij v angleškem jeziku, slovnica pa je tudi vsebinsko primerljiva z drugimi (češkimi, angleškimi), ki so služili kot model. Posledično je tudi vključena v splošno dostopni paket korpusov z besednimi skicami, ki jih ponuja orodje

Sketch Engine na strežniku http://www.sketchengine.co.uk/. Tretja, kasnejša verzija SBS (Krek 2012; Kosem et al. 2013; Kosem et al. 2013a) je bila izdelana za namen avtomatskega luščenja skladenjskih struktur, kolokacij in zgledov za podatkovno bazo kolokacijskega slovarja za slovenski jezik (Gantar et al. 2015). Ta verzija vsebuje bistveno več relacij, ker mora zagotoviti natančnejše luščenje kolokacij (in zgledov) po posameznih besednih vrstah ter znotraj besednih vrst po sklonih in drugih lastnostih, ki so uporabljene v SBS. Kot taka je manj primerna za ročno pregledovanje v leksikografskem procesu (Klosa 2013), zagotavlja pa vnaprej pripravljene kolokacijske podatke, kar pomembno skrajša čas, potreben za izdelavo slovarja (Kosem et al. 2013a).

3 SKLEP

Glede na anketo, ki je bila opravljena v okviru evropskega projekta COST *European Network of e-Lexicography* (Krek et al. 2015), je Sketch Engine med najbolj priljubljenimi orodji, ki jih v okviru leksikografskega procesa uporabljajo evropski leksikografi. Kot poleg tega kaže analiza avtomatizacije procesov in strojnega luščenja leksikografsko relevantnih podatkov iz besedilnih korpusov v okviru istega projekta (Tiberius et al. 2015), je vključevanje postopkov strojnega procesiranja naravnih jezikov v leksikografsko delo med leksikografi že sedaj uveljavljen proces, ki se bo v prihodnosti po pričakovanjih še intenziviral, kar pomeni, da se bo avtomatizaciji leksikografskih postopkov težko izognil katerikoli večji leksikografski projekt. Opisana slovnica besednih skic je (bila) namenjena podpori »ročnemu« leksikografskemu delu pri izdelavi Leksikalne baze za slovenščino, torej za sprotno (ročno) prenašanje podatkov iz korpusa v slovarsko bazo. Glede na predvidevanje, da bo velik del procesov pri izdelavi slovarjev v prihodnje vsaj na začetku izpeljan avtomatsko (Krek et al. 2013b), lahko predpostavimo, da bodo leksikografi potrebovali dva komplementarna pogleda na korpusne podatke. Prvi pogled, ki ga tu ne opisujemo, bo prilagojen avtomatskemu izvozu podatkov iz korpusa na začetku procesa, drugi pa robustnemu pogledu na celoto, ki zagotovi hiter skladenjsko-kolokacijski pregled okolice izbrane leme. Opisana slovnica je prilagojena temu drugemu namenu, ki ima znotraj leksikografskega procesa še vedno svoje legitimno mesto kljub zaželeni čim večji avtomatizaciji vseh delov procesa.

VI
(Iz)govorjeno v slovarju

Izgovor v slovarju sodobnega slovenskega jezika

Peter Jurgec

Abstract
This paper presents pronunciation guidelines for the new dictionary of contemporary Slovenian. It addresses three key issues: (i) the amount of information to be included; (ii) the method of transcription to be used; and (iii) the variety of Slovenian to be chosen as standard. These issues are affected by two crucial factors: the user-friendliness of the dictionary and timing considerations. Given this, I propose a minor modification to the traditional Slovenian transcription.

Ključne besede: izgovor, slovarji, fonetika, fonologija, naglas, morfologija, oblikoslovje

Keywords: pronunciation, dictionaries, phonetics, phonology, stress, prosody, morphology

1 UVOD

V slovenski slovaropisni tradiciji v tem in prejšnjem stoletju je imel izgovor eno izmed osrednjih vlog. Tako Slovar slovenskega knjižnega jezika (SSKJ) kot Slovenski pravopis 2001 (SP 2001) imata obsežna uvoda, ki se podrobno ukvarjata tudi z izgovornimi, fonetičnimi, fonološkimi in morfonološkimi vprašanji. Ta prispevek nakaže izgovorne smernice v novem slovarju sodobnega slovenskega jezika – omenja le ključne izzive pri načrtovanju novega slovarja in nakaže najpomembnejše smernice pri iskanju odgovorov.

Glavni smernici sta, prvič, prijaznost do uporabnika, torej enostavnost, jasnost in uporabnost, ter drugič, možnost realizacije v jasno določenem in omejenem časovnem okviru. Ti smernici narekujeta rešitve glede obsežnosti izgovornih informacij v slovarju (razdelek 2) in načina označevanja (razdelek 3). Rešitve so umeščene v širši okvir stanja slovenskih glasoslovnih raziskav (razdelek 4).

2 OZADJE ALI KOLIKO OZNAČITI V SLOVARJU

Izgovor je imel osrednjo vlogo v dosedanjih slovarskih projektih in tudi v načrtu novega slovarja, ki nastaja na Inštitutu za slovenski jezik Frana Ramovša (Gliha Komac et al. 2015). Razlogov za to je več. Prvič, izgovarjava iz zapisa ni vedno predvidljiva. To ne velja samo za naglasno mesto, temveč tudi kvaliteto naglašenega samoglasnika. Slovenščina ima samo 5 znakov za 8 ali 9 razlikovalnih samoglasnikov (Jurgec 2011). Tako *e* stoji namesto katerega koli nezadnjega sredinskega samoglasnika {e, ɛ, ə}. Drugič, naglas odvisnih oblik ni vedno predvidljiv iz osnovne oblike (tipično imenovalnik ednine za samostalnike, nedoločna oblika pridevnika v ednini moškega spola za pridevnike in nedoločnik za glagole). Pri samostalnikih je imenovalnik ednine pogosto izjemen v primerjavi z vsemi ostalimi oblikami.

Poglejmo si primer samostalniških iztočnic moškega spola. Izgovorno se samostalniki po obsežnosti precej razlikujejo. V (1) jih razvrščam v tri skupine: najenostavnejši tip, kjer je naglasno mesto nespremenljivo skozi paradigmo, enako velja tudi za kvaliteto samoglasnika in deloma tudi tonem, ki pozna predvidljive spremembe (kar v slovarjih običajno ni označeno). Srednje zapleteni primeri imajo premene v naglasnem mestu in kvaliteti samoglasnika (ali oboje). Ti primeri najbolje kažejo, da je koren v imenovalniku ednine pogosto izgovorno poseben v primerjavi z vsemi drugimi skloni. Če primerjamo različico s premičnim naglasom *zákon* ~ *zakóna*, je naglas na prvem zlogu korena samo v imenovalniku ednine, vsi drugi skloni pa imajo naglas na drugem zlogu. Podobno razmerje med

imenovalnikom ednine in ostalimi skloni je v precejšnjem delu samostalnikov moškega in srednjega spola.

(1) Primeri samostalnikov prve moške sklanjatve

	A. Enostavno	B. Srednje	C. Zapleteno
Zapis	gozdar	zakon	gozd
Zapis z naglasom	gozdár gozdárja	zákon zakóna zákon zákona	gòzd gôzda gózd gózda gozdôvi gôzdi gózdi
SSKJ	gozdár -ja (á)	zákon -óna *tudi* -a (á ǫ́; á)	gòzd gôzda *tudi* gózd -a *mn.* gozdóvi *tudi* gôzdi *tudi* gózdi (ȍ, ó; ǫ́)
SP 2001	gozdár -ja (á)	zákon -óna *tudi* zákon -a (á ǫ́; á)	gòzd gôzda *tudi* gózd -a *mn.* gozdôvi *tudi* gôzdi *tudi* gózdi -ov (ȍ, ó; ǫ́)

Fonološka izjemnost imenovalnika ni vezana samo na naglas, temveč tudi na več drugih spremenljivk. Primer *gozd* v (1) ima premeno v trajanju samoglasnika in zvenečnosti (ki pa nista razlikovalna), drugi samostalniki pa poznajo še premene v končnem korenskem samoglasniku (tip *Marko* ~ *Marka*) ali soglasniku (tip *carj* ~ *carja*). Če povzamem, s slovarskega stališča je izziv morfonološki: imenovalnik ednine je pogosto najbolj izjemna oblika. Zadnja skupina besed v (1) predstavlja iztočnice, ki niso izjemne samo v imenovalniku ednine, temveč tudi v drugih sklonih. Vzorčni primer *gozd* ima izjemne množinske oblike.

Moj namen na tem mestu ni razpravljati, katere oblikoslovne informacije naj bi bile v slovarju (o tem gl. K. Dobrovoljc 2015), ampak opoziriti, da oblikoslovne značilnosti slovenščine vplivajo že na zapis iztočnice. Če omejim razpravo na naglasna znamenja, imajo ta v slovenski glasoslovni tradiciji štirikratno vlogo: poleg mesta naglasa označujejo še kvaliteto naglašenega samoglasnika, kvantiteto in tonem. Kljub temu da je uporaba slovarjev eden izmed ciljev srednješolskega izobraževanja, uporabniško prepoznavanje naglasnih znamenj še ni bilo posebej raziskano (o tem gl. Rozman et al. 2015 in Arhar Holdt 2015). Naglasna znamenja zapisana v zaporedju v (1) so za uporabnika nejezikoslovca nedvomno precejšen izziv, posebej še, ker ima lahko isti simbol različne pomene, glede na to, kje v geselskem članku se pojavlja. Sestavljeni simbol *á* tako lahko pomeni naglašeni [a] v besedi kot *párk*, del pisne podobe (vendar ne naglasa) kot v besedi *Dvořák* ali akutirani tonemski naglas, kakor oklepajno pri geslu *gozdár* -ja (á), tako kot v (1).

Ali je povprečni uporabnik res dovolj izurjen, da lahko loči med različnimi pomeni istega simbola? Mar ne bi bilo bolj smiselno izbrati le tistih informacij, ki so za

uporabnika ključne in jih predstaviti kar se da enostavno? Sem vsekakor spadata naglasno mesto in kvaliteta samoglasnika, manj pa kvantiteta in tonem.

Pregledal sem tri ključna vprašanja in fonološke značilnosti slovenščine, ki narekujejo, kako naj bi bila izgovarjava označena v slovarju slovenskega jezika.

3 PREDLOG ALI KAKO OZNAČITI V SLOVARJU

Naslednje vprašanje je, kako označiti izgovarjavo, da bo do uporabnika prijazna, jasna in hkrati dovolj natančna. Na tem mestu predlagam kombinacijo obstoječe slovenske transkripcije v kombinaciji z nesimbolnim zapisom, ki bo rešil težave z dvoumnostjo naglasnih znamenj.

Velika večina besed v slovenščini ima pisno podobo brez diaktritičnih znamenj, medtem ko ima manjši del prevzeta diaktrična znamenja iz drugih jezikov. Dobro znani slovaropisni izziv je, da se ista diakritična znamenja lahko uporabljajo tudi za naglas. SSKJ in SP 2001 razlike med obema vrstama iztočnic nista jasno označevala, tako da ni bilo vedno povsem jasno, ali je naglasno znamenje del pisne podobe ali ne. Slovar novejšega besedja slovenskega jezika (Bizjak Končar et al. 2014) je bil tu nekoliko bolj konsistenten, saj naglasa ni označeval pri iztočnicah, ki so imele dodan izgovor, pri vseh ostalih pa je bil naglas označen na iztočnicah. Kot enostavnejša in še jasnejša rešitev bi bilo označevanje naglasa s kakšnim drugim simbolom, ki ni del pisne podobe nobene iztočnice. Taka rešitev bi bila lahko podčrtava črke, ki odgovarja naglašenemu glasu (tipično samoglasniku). Težava pri tem je, da se tipografska pravila kar naprej spreminjajo in je težko predvideti, kaj bo mogoč del pisne podobe vseh prihodnjih iztočnic. Alternativna možnost bi bila označevanje izgovora kako drugače, s spremembo nabora znakov (fonta) ali barvo tiska. Ker je načrtovani slovar primarno spletne narave, predlagam slednjo rešitev (2). Dvobarvni tisk bo mogoče prenesti tudi v knjižno izdajo. Izgovor osnovne oblike je potem podan v celoti v oglatih oklepajih, pri čemer se uporablja ortografska podoba z tradicionalnimi slovenskimi oznakami za netonemski naglas. Tako se vzpostavi popolna ločitev pisne podobe iztočnice od izgovora.

(2) Predlog za označevanje naglasa

a. Iztočnice brez diakritičnih znamenj, ki so del pisne podobe
   gozdar [gozdár]
   zakon [zákon]
   gozd [gòzd], [gózd]

b. Iztočnice z diakritičnimi znamenji, ki so del pisne podobe
à la carte [a la kárt]
skijöring [skjêring]
fin de siècle [fen də sjékəl]

Izgovor je zapisan kar se da blizu pisavi, pri čemer niso upoštevana vsa fonološka pravila, v skladu z dosedanjo prakso v slovenskem slovaropisju. Označen je izgovor polglasnika (3-a) in v-jevski izgovor črk *l* (b) ali *u*, in sicer s simbolom [w], ki je splošnemu uporabniku blizu. Izgovor črke *v* načeloma ni označen.

(3)

a. Polglasnik
pes [pə̀s]
dež [də̀ž]
sen [sə̀n]

b. Premene l-ja
pol [pôw]
fižol [fižôw]
žival [živáw]
pav [páv]

Velika večina besed ima v slovenščini eno samo, stalno naglasno mesto. Slovenska glasoslovna literatura pa pozna tudi breznaglasnice, ki niso naglašene, razen če so izgovorjene samostojno, in večnaglasnice, ki imajo dva ali več besednih naglasov (Toporišič 1969, 1976/2000; Jurgec 2007). Splošna fonološka teorija podobnih izrazov ne pozna, uporabljata pa se izraza primarni in sekundarni naglas.

Označevanje breznaglasnic doslej ni bilo sistematično urejeno in naglas je včasih označen (npr. *zaradi* v SSKJ in SP 2001), drugič ne (*se*). V slovarju je vsekakor treba označiti razliko med vedno in včasih naglašenimi iztočnicami. Pričujoči predlog omogoča takšno razlikovanje: medtem ko imajo naglašene iztočnice označen naglas z drugačno barvo, breznaglasnice take označbe nimajo. Oklepajni izgovor obeh vrst gesel pa vsebuje informacijo o naglasu (4).

(4) Breznaglasnice
tudi [túdi]
zaradi [zarádi]
ne [nè]
se [sè]
že [žé]
toda [tóda]

Kar pa se tiče večnaglasnic, slovaropisna praksa izkazuje precejšnjo neenotnost pri določanju naglasa na nekončnih korenih in na predponah. Potrebne bodo še nadaljnje raziskave, ki presegajo okvirje tega slovaropisnega projekta. Pragmatična rešitev je označevanje samo enega, torej primarnega naglasa, kar je prikazano v (5).

(5) Večnaglasnice

	SSKJ	SP 2001	SSSJ (predlog)
mikrometer	mikrométer -tra (ẹ̄)	mikrométer -tra (ẹ́)	mikrometer [mikrométər]
gigameter	gígaméter -tra (ȋ-ẹ̄)	gígaméter -tra (ȋẹ́)	gigameter [gigamétər]
elektrometer	elektrométer -tra (ẹ̄)	elektrométer -tra (ẹ́)	elektrometer [elektrométər]
elektroenergija	eléktroenergíja -e (ệ-ȋ)	eléktroenergíja -e (ẹ́ȋ)	elektroenergija [elektroenergíja]
telekamera	télekámera -e (ệ-â)	télekámera -e (ẹ́â)	telekamera [telekámera]
telekomunikacija	telekomunikácija -e (á)	telekomunikácija -e (á)	telekomunikacija [telekomunikácija]

Glavni vodili pri omenjenih rešitvah sta prijaznost do uporabnika in neodvisnost od dodatnih raziskav. Prijaznost do uporabnika se kaže v intuitivni uporabi barve za naglasno mesto in uporabi obstoječih naglasnih znamenj za izgovor, z izjemo polsamoglasniškega izgovora črke *l*, kjer prevladata splošna prepoznavnost *w* kot simbola za tak glas, prevzetega iz angleške ortografije. Pričujoče rešitve ravno tako ne zahtevajo dodatnih raziskav glede sekundarnega naglasa v slovenščini ali izgovarjave polglasnika.

4 RAZPRAVA ALI KAKŠEN IZGOVOR IZBRATI

Zadnje vprašanje tega prispevka je izbira izgovora. Vprašanje bom obravnaval v dveh delih. Prvič, katera so ključna neobdelana vprašanja, ki so izziv za slovarske projekte. Drugič, kaj je reprezentativni izgovor slovenščine, ki naj bi bil v slovarju.

Vsak večji slovaropisni projekt v slovenistični tradiciji razkriva, kako slabo je raziskana fonologija slovenščine. S slovaropisnega vidika se glavna vprašanja dotikajo naglašenih samoglasnikov in naglasa. V predhodnem razdelku sem že omenil

nekaj takih vprašanj, tukaj pa bom izpostavil nekaj možnih rešitev. Najbolj pereč je naglašenost nekaterih tvorjenk (zloženk, sestavljenk in sklopov), ki niso sistematično urejene. Med rešitvami je precej neenotnosti, zlasti pri določevanju naglasnega mesta predpon in nekončnih korenov. Pregled gradiva je pokazal, da naglašenost zadnje sestavine zloženk in sklopov načeloma ni problematična, drugod pa je več variantnosti (Jošt v pripravi). Statistična analiza kaže, da je naglašenost predpone odvisna tudi od tega, ali je naglašen prvi zlog osnove. To v celoti nasprotuje slovenističnim ugotovitvam o večnaglasnicah, je pa v skladu z ugotovitvami splošnega jezikoslovja (Liberman in Prince 1977; Halle in Vergnaud 1987; Burzio 1994; Hayes 1995; Kager 2007; van der Hulst 2010).

Zrcalni problem je naglašenost predlogov, veznikov, členkov, zaimkov in glagola *biti*; gre za skupino besed, ki je v slovenščini znana kot breznaglasnice. Težava slovenistične teorije o breznaglasnicah (Toporišič 1976/2000, 2001 in drugod) je, da ne ločuje med besedo v morfološkem in fonološkem smislu (torej prozodično besedo). Če so besede definirane v razmerju do svojih fonoloških lastnosti, kamor spadata besedni naglas in premene na besednih mejah, potem breznaglasnice ne predstavljajo večje težave. V Jurgec (2007) predlagam natanko to: breznaglasnice v ustreznem besedilnem kontekstu niso samostojne prozodične besede, postanejo pa lahko, če so izgovorjene samostojno, kar je v skladu z ugotovitvami splošnega jezikoslovja (npr. Selkirk 1980; Nespor in Vogel 1986; Peperkamp 1997; Pierrehumbert 2003; Itô in Mester 2003, 2009). Z naglasom povezano vprašanje je tudi kvaliteta samoglasnikov v nekaterih prevzetih besedah, tip *maraton*, *Washington*, *Maribor*, *Bangladeš*, *Kosovel*, *Izrael*. V Jurgec (2010) te primere razložim z naglasom na ravni stopice, ki prej za slovenščino ni bila predlagana. Pri kratičnih imenih gre ponavadi za dva naglasa: pomembnejši je na zadnjem samoglasniku, stranski pa na prvem, tip *cede*, *PTT*, *wc*, *ZZZS*. V Bizjak Končar et al. (2014) so nekateri izmed teh primerov rešeni z ozkimi samoglasniki na prvem zlogu in naglasu na zadnjem zlogu.

Argument, zakaj sekundarnega naglasa ni potrebno označevati, je takle. Prvič, sekundardni naglas v prevzetih besedah in kraticah je predvidljiv (kar pa ne velja za zloženke). V prevzetih besedah je na vsakem drugem zlogu, šteto od naglašenega zloga na obe strani; v kraticah je sekundarni naglas na prvem zlogu. Drugič, da je beseda prevzeta ali kratica, je očitno iz drugih informacij v geselskem članku.

Tudi polglasnik je eden izmed problematičnih vprašanj slovenske fonologije. Medtem ko je distribucija, ki temelji na slovaropisni tradiciji in ugotovitvah zgodovinskega jezikoslovja dobro opisana (Toporišič 1976/2000), ti podatki pogosto niso v skladu z dejanskim govorom. Eden izmed takih pojavov je premena polglasnika s sprednjim samoglasnikom v izpeljankah. Polglasnik v položaju pred *r* tozadevne premene ne pozna. Več podatkov je v (6).

(6) Polglasniške premene v izpeljankah

Osnova		Izpeljanka	
pə̯s	‘pes’	pe̯sjak	‘pesjak’
kə̯s	‘kes’	ke̯sanje	‘kesanje’
mə̯gla	‘megla’	me̯glenost	‘meglenost’
tə̯ma	‘tema’	te̯mačən	‘temačen’
mesə̯c	‘mesec’	mese̯čina	‘mesečina’

Tovrstni vzorci, razen izjemoma, v sodobni slovenistiki sploh še niso bili zaznani (Šeruga Prek in Antončič 2003; Jurgec 2007, 2011). Ti primeri sicer ne predstavljajo večjih težav za slovar, saj bodo imele osnove in izpeljanke v (6) samostojne geselske članke. Polglasnik bo označen samo pri prvi skupini.

Neopisanost fonologije slovenščine najbolj razgali dejstvo, da v slovenistiki ni enotnega mne-nja niti o tem, koliko samoglasnikov slovenščina zares ima in ali je samoglasniška kvantiteta sploh razločevalna. Tu pa preidemo na drugo vprašanje, namreč katere različice slovenščine bi morale biti v slovarju vključene in katere ne. S stališča izgovora sta se v slovenistični literaturi pojavaljala vsaj dva modela: (i) posplošeni slovenski govor, ki morda nima rojenih govorcev (Toporišič 1976/2000; Lenček 1981; Herrity 2000; Tivadar 2004a, 2007, 2010; Tivadar in Tivadar 2015) ali (ii) govor izobraženih govorcev osrednjih narečij v formalnih govornih položajih (Bezlaj 1939; Toporišič 1975; Srebot Rejec 1988, 1998; Šuštaršič et al. 1995, 1999; Petek et al. 1996; Jurgec 2005, 2006a, b; Tivadar 2004b, 2008, 2010). Za daljšo razpravo o tem gl. Jurgec (2011).

Odločitev med zgoraj omenjenima možnostma se v tem trenutku zdi bolj vprašanje ideologije kot znanosti, ima pa posledico za novi slovar. Ena možnost bi bilo iskanje novih rešitev, kar bi ob dodatno potrebnih raziskavah verjetno vzelo kakšno desetletje ali več. Druga možnost je opisati slovar z obstoječim védenjem in ob zavedanju pomanjkljivosti. Slednja rešitev je trenutno edina realno izvedljiva (za alternativno, veliko bolj ambiciozno, a glede na trenutno stanje fonoloških raziskav težko izvedljivo rešitev gl. Gliha Komac et al. 2015). V zvezi z izgovorom to pomeni, da slovar ne bo prinašal novih, širših, metodološko in teoretično podkovanih fonoloških raziskav, ampak bo prevzel osrednje elemente obstoječih dognanj z zavedanjem omejitev.

Takšna omejitev bo neoznačevanje tonemskega naglaševanja, ki je že od nekdaj zanimala samo ozek krog strokovnjakov, neznanstvenemu uporabniku pa je v celoti nezanimiva. Tovrstne informacije bodo še vedno na voljo v obstoječih

slovarjih (SSKJ in SP 2001). Večnaglasnice bodo imele označene samo zadnji naglas, kot v (5). Izgovorne posebnosti in variantnost pa se bodo preverjale anketno. Pri tem bodo izbrane ključne besede, ki bodo ustrezno onaglašene s pomočjo množičenja (o množičenju Fišer in Čibej 2015 ter Fišer et al. 2015).

5 SKLEP

Članek predstavi glavna izhodišča za določevanje naglasa v slovarju sodobnega slovenskega jezika. Slovar bo vseboval vse za uporabnika nepogrešljive informacije, kakršno je naglasno mesto, kvaliteta naglašenega samoglasnika, polglasniški izgovor in izgovor črke *l*. Večine problematičnih primerov v slovarju ne bo. Tako slovar ne bo vključeval predvidljivega sekundarnega naglasa v prevzetih besedah tipa *Kosovel*, *Sarajevo*, sekundarnega naglasa v kratičnih imenih in sekundardnega naglasa v nekaterih tvorjenkah. Druge problematične primere (naglasno mesto in kvaliteta naglašenega samoglasnika) ter variantnost bomo preverjali s spletnim anketiranjem. Rezultat bo do uporabnika prijazen in ga bo mogoče realizirati v krajšem časovnem okviru.

Govorjeni proti pisnemu ali katera leksika je »tipično govorjena«

Darinka Verdonik

Abstract
This paper aims to resolve two key issues concerning the usage of the GOS speech corpus as a data source for lexicographic descriptions: (i) the words which are especially common in the GOS corpus; and (ii) whether it is appropriate to mark such words or some of their usages as »spoken«. Based on the key-word corpus analysis performed, seven different groups of words were defined as being common in the GOS corpus: non-verbal expressions; discourse markers; stylistically marked lexis; deixes; expressions for speech acts that are common in spoken usage; expressions that are »favoured« in spoken usage; and forms that are more favoured in spoken that in written usage. The discussions conclude with the argument that marking these words or some of their usages as »spoken« is too superficial an annotation because the factors which encourage their usage are connected with different levels of discourse, such as formality, spontaneity, common speech acts, etc. We suggest more specific register-stylistic marks, such as »non-formal«, »in conversation«, »expression of attitude«, »regionally specific usages«, etc.

Keywords: speech, speech corpus, spoken lexis

Ključne besede: govor, govorni korpus, govorjena leksika

1 UVOD

V zadnjem desetletju in več lahko sledimo porastu govornih korpusov za številne jezike (gl. Verdonik et al. 2013). Pogosto so bili uporabljeni v različnih diskurznih in pragmatičnih študijah (npr. Aijmer 1996; Martinez 2011; pri nas npr. Verdonik in Kačič 2012), tisti, ki so bili spodbujeni z govornimi tehnologijami, so pomenili ključen dejavnik predvsem v razvoju razpoznavanja tekočega govora (Žgank et al. 2008; Golik et al. 2013), svojo nišo išče t. i. korpusno jezikoslovje govora (Adolphs in Carter 2013), zlasti v angleškem jezikoslovju pa je že več desetletij aktualno tudi vprašanje, kako ravnati z govornimi korpusi kot leksikografskimi viri (Svartvik 1992; Moon 1998; De Cock 2002). Tudi v zadnjem desetletju zasledimo vsaj dve vidnejši, eksplicitni obravnavi tega vprašanja v mednarodnih leksikografskih publikacijah.

Trap-Jensen (2004) na podlagi izkušenj z danskim slovarjem poudarja, da so iz govornega dela korpusa, uporabljenega za izdelavo slovarja, dobili velik del informacij, vključenih v slovar. Med leksiko, ki so jo označili kot rabljeno »zlasti v govorjenem jeziku«, naštevajo medmete in onomatopeje[1] (*mhm, aha, eee, kikiriki*), diskurzne označevalce (*veš, mislim, pol*), pragmatične fraze (*oprostite, ti povem, kaj ne poveš, daj no*), deiktične zaimke in prislove (*ta, oni*), kletvice (*presneto, hudiča*) ter slengovske in nestandardne izraze (*ali pa kaj takega, daj nehaj*), vključno z regionalizmi in narečno leksiko, ki je razširjena čez lokalne meje svojega izvora. Ob naštetih skupinah od deiktičnih izrazov naprej že avtor sam izrazi pomislek, ali jih je bolje označiti kot »neformalne« ali kot »značilne za govorjeno rabo«, pri regionalizmih pa se sprašuje, ali jih ne bi bilo bolje označiti glede na regionalno razširjenost. Ob naštetem ni zanemarljivo, da navedeni danski slovar temelji na korpusu obsega (samo) 40 mio. besed, od tega 8 mio. v govorjeni rabi in 32 mio. v pisni. Za slednje je korpus takega obsega z današnjega gledišča (prim. Gigafida – 1 milijarda besed) zelo majhen. Trap-Jensen (2004) argumentacija, da brez 315 eksplicitnih navedkov, ki so jih vključili na podlagi govorjenega gradiva v njihovem korpusu, izdelani danski slovar ne bi bil popoln, je zaradi tega lahko nekoliko manj prepričljiva.

Siepmann (2015), nasprotno, nima težav s pomanjkanjem gradiva po današnjih standardih: govorni korpus, ki ga uporablja, CRFC – Referenčni korpus sodobne francoščine (Siepmann in Bürgel 2015), obsega 155 mio. besed v govorjeni rabi, podatke v njem pa primerja z opisi v 10 različnih obstoječih in najširše uporabljanih eno- in dvojezičnih slovarjih francoskega jezika. Ob tem dokazuje, da nobeden od obstoječih francoskih slovarjev za izbrane analizirane

1 Navajamo sorodne slovenske zglede, ki smo jih izbrali na podlagi danskih zgledov v Trap-Jensen (2004). Dobesedno prevajanje danskih zgledov ne bi imelo smiselnega učinka, saj so v različnih jezikih za opravljanje podobnih pragmatičnih funkcij rabljeni različni izrazi. Zgledi so navedeni izključno z namenom, da si bralec bolje predstavlja, kaj v danskem slovarju razumejo pod navedenimi poimenovanji.

štiri enobesedne in tri večbesedne leksikalne enote ne opisuje nekaterih najbolj pogostih, vsakdanjih leksikalnih vzorcev in kolokacij, ki jih najdemo le v govornih korpusih, v pisnih, še tako velikih, pa ne. Čeprav Siepmann (2015) meni, da bi slovaropisci te značilnosti govorjene rabe lahko opisali na podlagi lastne jezikovne intuicije, pa je po drugi strani jasno, da je dodajanje opisov pomenov ali zgledov na podlagi lastne jezikovne intuicije v slovaropisju (vedno bolj) nesprejemljivo. Prav zaradi tega pa ima Siepmannov članek še večjo težo, s tem ko dokazuje, da lahko (1) srednje veliki korpusi, kot je CRFC, osvetlijo leksikalne vzorce in kolokacije, o katerih iz še tako velikih pisnih korpusov ne izvemo nič, in da (2) imajo t. i. govorjene (angl. *colloquial*) leksikalne enote svoje lastne pomenske odtenke, frazeologijo in pragmatiko ter nikakor niso slogovno manj vredni nadomestki za bolj formalne enote (povzeto iz Siepmann 2015). Tovrstna leksika in vzorci njene rabe so po mnenju Siepmana izredno globoko zakoreninjeni v rutinah vsakdanjega življenja, hkrati pa v veliki meri še neopisani.

Oba predstavljena članka skušata dokazati, da se v govorjeni rabi z leksikalnega vidika pojavljajo rabe, ki bi sodile v slovarski opis, a tam pogosto manjkajo. Medtem ko je Trap-Jensen (2004) nakazoval, da lahko vrsto leksikalnih enot, od neverbalnih prek deiktičnih do nestandardnih in regionalno značilnih izrazov, označimo kot tipično govorjene same po sebi, se je Siepmann (2015) osredotočal predvsem na nove pomenske in funkcijske odtenke v govorjeni rabi ter tipično govorjene kolokacije za nekaj izbranih, pogosto rabljenih izrazov.

V tem prispevku bomo vprašanje, kaj je tipično za govorjeno rabo, dodatno raziskali. Ob korpusih, ki so na voljo za sodobno slovenščino, nas bo zanimalo, katera leksika se v govorjeni rabi pojavlja bolj pogosto kot v pisni, kaj so razlogi za to in ali jo smemo samo zaradi njene večje frekventnosti v slovarju označiti kot »tipično za govorjeno rabo«. Pri tem se bomo omejili na enobesedne leksikalne enote.

2 JEZIK V GOVORJENI IN PISNI RABI

Raziskave, ki primerjajo jezik v obeh kanalih, govornem in pisnem, rade izpostavljajo razlike. Vse od 20-ih let 20. stoletja so se vrstile analize (pretežno na angleščini), ki so dokazovale kvantitativne razlike v dolžini besed, povedi, besedil, variabilnosti slovarja, kasneje, v drugi polovici 20. stoletja, tudi razlike v težavnosti, abstraktnosti, zgoščenosti in razumljivosti besedil v enem in drugem kanalu ter v distribuciji besed v njih. Brez dvoma prehod iz govornega v pisni kanal in obratno spremljajo tudi spremembe v vzorcih leksikalne izbire, povzame pregled teh raziskav Akinnaso (1982: 103).

Tudi v 80-ih letih in vse do danes se nadaljujejo predvsem kvantitativne primerjave govorjenih in pisnih besedil. Avtor številnih je Biber (npr. 2009; 2012; et al.), ki s korpusnojezikoslovnim pristopom dokazuje, da so znatne razlike v leksikalnih vzorcih med neformalnim pogovorom in akademskim/znanstvenim pisanjem. Med drugim opozarja, da splošni opis leksikalno-skladenjskih vzorcev, kot ga najdemo v Longmanovi slovnici govorjene in pisne angleščine (Biber et al. 1999), vključuje nekatere značilnosti, ki jih ne najdemo niti v neformalnem pogovoru na eni strani niti v znanstvenih besedilih na drugi strani, ob tem pa ne vključuje nekaterih pomembnih značilnosti, ki jih najdemo v enem ali drugem tipu diskurza (Biber 2012).

Toda faktorjev, ki vplivajo na razlike v jezikovni rabi v neformalnem pogovoru na eni in znanstvenih besedilih na drugi strani, je veliko, ne samo različen kanal. To bi lahko vrglo nekaj dvoma na zaključke Biberjevih raziskav. Vendar tudi raziskave z manj diametralno nasprotnim gradivom dokazujejo pomembne razlike. Chafe in Danielewich (1987) primerjata jezikovne vzorce v angleščini v različnih žanrih in kanalih komunikacije v akademskem okolju. V raziskavo zajameta vedno istih 20 govorcev, zastopani žanri pa so: predavanja, pogovori v univerzitetnem kampusu, pisma in znanstveni članki. Na ravni leksike ugotovita, da je tako v predavanjih kot v pogovorih (torej obeh žanrih govorjene rabe) slovar relativno omejen v primerjavi s pisnima žanroma, uporaba omejevalcev (angl. *hedges*) je večja in referenčna neeksplicitnost pogostejša. Heinrichsen in Allwood (2005) primerjata neformalne pogovore in sociolingvistične intervjuje s časopisnimi besedili v danščini in švedščini ter trdita, da je več podobnosti med govorjeno rabo v danščini in švedščini (torej dveh sorodnih, a različnih jezikih) kot med pisno in govorjeno rabo v istem jeziku, danskem ali švedskem. Razlike opazujeta v razpršenosti slovarja, natančnosti izražanja, distribuciji besed in besednih vrst itd.

Z razmahom raziskav diskurznih označevalcev se pojavijo tudi razprave, ki utemeljujejo, da lahko tovrstne vloge leksikalnih enot najdemo predvsem v govorjeni rabi in da predstavljajo na leksikalni ravni eno najbolj izrazitih razlik v razmerju do pisne rabe. Adolphs in Carter (2003) tako na primeru angleške besede *like* dokazujeta, da nam uporaba govornega korpusa pomaga odkriti značilnosti govorjene rabe, ki pred tem niso bile sistematično identificirane. Podobno dokazuje Siepmann (2015) za sedem izbranih izrazov v francoščini, kot smo že predstavili v uvodu.

Čeprav od Akinnasovega (1982) pregleda raziskav, ki primerjajo govorjeno in pisno rabo, mineva več kot trideset let, so njegovi zaključki še vedno aktualni. Ozko kvantitativna primerjava leksikalnih (in skladenjskih) enot lahko da popačeno sliko, trdi, in argumentira, da je treba razlike opazovati iz širšega, diskurznega stališča, saj lahko nanje vpliva veliko faktorjev, ne samo govorjenost

oz. pisnost: formalnost proti neformalnosti, tema pogovora, narava komunikacijskih ciljev itd. Govorjena in pisna raba jezika sta v mnogo pogledih komplementarni, zatrjuje Akinnaso: nekatere komunikacijske potrebe večinoma zadovoljujemo skozi govorjeno rabo, druge prevladujoče skozi pisno (1982). Te funkcijske razlike vplivajo na leksikalne, skladenjske in semantično-pragmatične lastnosti besedil v enem in drugem načinu jezikovne rabe. Ob tem, da predstavljene raziskave brez dvoma dokazujejo opazne razlike med obema načinoma jezikovne rabe, je zato treba temeljito premisliti, kaj so najpomembnejši faktorji, ki v resnici vplivajo na to.

3 GRADIVO IN METODA

Za primerjavo razlik v leksiki med govorjeno in pisno rabo smo uporabili korpusa Gos (Verdonik et al. 2013) in Gigafida (Logar Berginc et al. 2012). Oba sta dostopna tudi prek orodja NoSketchEngine (Erjavec 2013; Kilgariff et al. 2014), kjer so mogoče tudi nekatere statistične primerjave besedil v njih.

Analiza ključnih besed (Scott 1997) je dobro znana korpusna metoda, po kateri naredimo seznam leksike, ki po pogostosti rabe v analiziranem korpusu najbolj odstopa od rabe v referenčnem korpusu, iz česar sklepamo, da je za analizirani korpus ključna – tipična. Nas zanima, katera leksika je ključna za korpus govorjene slovenščine Gos, iz česar bi lahko sklepali, da je tipična za govorjeno rabo. Gigafida nam pri tem služi kot referenčni korpus. Za vsako različnico izračunamo, kolikokrat se pojavi na 1 milijon pojavnic v analiziranem (Gos) in referenčnem korpusu (Gigafida). Da se izognemo matematičnemu problemu, kadar se katera od različnic v enem od korpusov ne pojavi niti enkrat, prištejemo obema vrednostma parameter n (n = 100), nato pa delimo število pojavitev v analiziranem korpusu s številom pojavitev v referenčnem korpusu in rangiramo različnice glede na dobljeni faktor (Lexical Computing Ltd. 2013).

Ker nas zanima leksika z vidika slovarske obravnave, ta pa poteka znotraj slovarskih iztočnic, smo izračune delali na lemah, ne pa na besednih oblikah. Lematizacija obeh korpusov je bila avtomatska, zato smo pri tem zajeli tudi napake, ki nastanejo pri lematizaciji. Naslednji korak po izdelavi prvih seznamov ključnih lem je tako bil popravljanje napačno lematiziranih enot, ki so se pojavljale na prvih seznamih.

Ko smo dobili sezname ključnih lem, na katerih ni bilo več očitno napačno lematiziranih enot, smo opravili ročno analizo rezultatov. Za vsako od prvih 100 ključnih lem smo pregledali konkordance v korpusu Gos, po potrebi primerjali s konkordancami v korpusu Gigafida, iskali morebitne izstopajoče kolokacije,

pregledovali vrste rab in jih primerjali z rabo najbližjih sopomenk. Nazadnje smo skušali med prvimi 100 ključnimi lemami definirati skupine, ki so povezane s skupnimi vzroki za večjo frekventnost v korpusu Gos.

Nadaljevanje seznama ključnih lem od 100 naprej, s poudarkom na prvi polovici seznama, smo samo pregledali ter iskali morebitne dodatne, med prvimi 100 ključnimi lemami neopažene, a vseeno pogosto izstopajoče skupine in vzorce.

4 REZULTATI

Definiranje skupin lem, ki so povezane s skupnim vzrokom za večjo frekventnost v korpusu Gos, ni mogoče brez nekaj posploševanja. Rabe posamezne leme so pogosto različne, enako razlogi zanje. Upoštevali smo predvsem tisto, kar je najbolj izrazito in pogosto. Nekatere leme smo tako uvrstili v več skupin.

4.1 Neverbalni izrazi

Na vrhu seznama za govorjeno rabo ključnih lem se pojavljajo nekateri neverbalni elementi, tipični za spontano govorno produkcijo: to so *eee, eem, mmm, aaa, nnn*, pogosto označeni kot označevalci vrzeli, lahko pa jim pripišemo tudi pomembne metadiskurzne vloge (npr. pri menjavanju vlog). Ti izrazi se v pisni rabi praviloma ne pojavljajo in zanje niti še ni splošno sprejete standardizacije zapisa – ta se šele poskuša vzpostaviti prek korpusa Gos. Tudi če najdemo iste nize v pisnem korpusu Gigafida, imajo zato praviloma drugačno vlogo in pomen: lahko so napaka v lematizaciji, kratica, v enaki vlogi kot v govorjeni rabi pa le izjemoma, pri premem govoru.

S slovarskega vidika se pri teh izrazih postavlja vprašanje njihovega mesta v slovarju: kot jezikovni elementi, ki niti niso jasno verbalizirani, so zelo mejne enote in jih do zdaj slovarji niso vključevali v svoje opise.

4.2 Diskurzni označevalci

Naslednja skupina lem, ki izstopajo po pogostosti v govorjeni rabi, so tiste, ki lahko v posameznih oblikah nastopajo tudi kot t. i. diskurzni označevalci. Na seznamu prvih 100 ključnih lem jih je precej: *mhm, aha, aja, veš/veste* (*vedeti*), *okej, v bistvu* (*bistvo*), *glej/te* (*gledati*), *poglejte* (*pogledati*), *zdaj, mislim* (*misliti*), *vidiš/te* (*videti*), *v redu* (*red*), *nič, dobro, jah*.

Za diskurzne označevalce (Verdonik 2007) je značilno, da ne prispevajo k propozicijski vsebini, ampak imajo prevladujoče različne metadiskurzne vloge: vzpostavljajo povezave z vsebino predhodnega ali/in prihodnjega diskurza, pomagajo vzpostavljati in vzdrževati odnos med sogovornikoma/i, izražajo odnos/držo govorca do propozicijske vsebine oz. pomagajo organizirati potek diskurza (menjavanje vlog, menjavanje tem, zaključevanje pogovora) – pri čemer niso specializirani, ampak lahko posamezen diskurzni označevalec opravlja več vlog hkrati, velikokrat pa obstaja tudi močna vez s predstavnim pomenom in enoumno razlikovanje med rabo v diskurznih vlogah in propozicijski vsebini ni mogoče. Diskurzni označevalci so bili najbolj eksplicitno opaženi prav v govorjeni rabi (Schiffrin 1987), kar pa ne pomeni, da jezikovnih sredstev z (delno) podobnimi funkcijami ni tudi v pisni rabi, kjer so znani predvsem kot metabesedilo (prim. Pisanski Peterlin 2011).

4.3 Slogovno zaznamovana leksika

Kot posebno skupino lem smo označili tiste, ki jih slovensko normativno jezikoslovje označuje kot tako ali drugače jezikovno slogovno zaznamovane – po Toporišiču (2000) so označene kot neknjižne zvrsti in interesne govorice. Čeprav bi morda pričakovali, da bodo seznam za govorjeno rabo ključnih lem prevladujoče sestavljale leme iz te skupine, se v resnici njihovo število giblje nekaj nad eno desetino vseh ključnih lem. Na seznamu prvih 100 ključnih lem je namreč 8 takih, ki so v slovenskih referenčnih slovarjih (pravopisni in SSKJ2) že same po sebi označene kot slogovno zaznamovane: *ful, glih, fajn, evo, pizda, cajt, izgledati, okej*. Še dodatnih 6 lem je takih, ki se v govornem korpusu pogosto rabijo v pomenih, ki so označeni kot taki: *pol* v pomenu 'potem', *rabiti* v pomenu 'potrebovati, biti treba'; *znati* v pomenu 'vedeti'; *drugače* v pomenu 'sicer/sicer pa'; *enkrat* v pomenu 'nekoč', moči v pomenu 'morati' (tip *še bomo mogl vadet*).

Med navedenimi lemami so tri take, ki so v Gosu rabljene prevladujoče (ne pa izključno!) pri mladih: *ful, pizda* in *okej*. Na nadaljnjem seznamu ključnih lem je takih še več, npr. *kul, kurec, kurba, fak*. Pri tem pa je treba upoštevati, da Gos kot referenčni korpus majhnega obsega ni zadosten vir za zanesljivo prepoznavanje prvin govorice mladih. Tako ta kot tudi druge t. i. interesne govorice so sicer pogosto predmet zanimanja, vseeno pa v slovenskem jeziku še nimamo ustrezno obsežnega in reprezentativnega gradiva za te zvrsti, zato je označevanje tovrstnih, tj. slengovskih in žargonskih prvin v slovarju lahko precej nezanesljivo in v veliki meri temelji na intuiciji.

Na seznamu ključnih lem od 100 naprej sledi še veliko takih, ki so v dosedanji slovenski slovaropisni praksi označeni za slogovno zaznamovane (npr. *kao, probati, ajde, valjda, ziher, ratati, pasati, kul, ovi, fora, zrihtati, fejst, mejl, špilati, tavžent,*

dec, zastopiti), vendar zaradi nizke pojavnosti o njihovi podrobnejši slogovno-zvrstni opredelitvi večinoma ni mogoče zanesljivo sklepati. Vseeno opozorimo vsaj na to, da se začne pojavljati tudi leksika, ki je v rabi regionalno precej ozko omejena, npr. *gučati, ovi, tipo, hambrt, vsikdar, gvišno*. Za manjši delež tovrstne leksike, ki se v Gosu dovolj pogosto pojavlja, bi se dalo zelo na splošno sklepati o njeni regionalni razširjenosti, vendar je velika težava z leksikografskega vidika pogosta napačna lematizacija, ki jo je treba v drugi verziji korpusa Gos kolikor mogoče odpraviti. Za večji del tovrstne leksike (npr. *farba, murka, ahtati, štuk, šimfati, štrihati, zalimati*) pa je Gos premajhen korpus, da bi lahko zanesljivo sklepali o njeni regionalni razširjenosti.

Kot nekoliko posebno, vseeno pa najbližjo temu sklopu označimo za govorjeno rabo povsem specifično lemo *ne_biti*, ki je pripisana pojavnicam, v katerih sta v govorjeni rabi dve lemi fonetično združeni v eno: *nem, nam, nevš, nev, nam, neb, neo, nem, nav, nevmo, nevte* ... Tukaj izgovorni elementi v korpusni strukturi prehajajo na raven lem; z vidika naše analize bi take strukture bolj ustrezno obravnavali po njihovih sestavnih elementih.

4.4 Deiktiki

Naslednja skupina za govorjeno rabo ključnih lem, ki smo jo definirali, so izrazi, s katerimi se orientiramo v prostoru in času ter imajo (tudi) deiktično funkcijo. Na seznamu prvih 100 najbolj ključnih lem je takih kar nekaj, pojavljajo se v parih: *pol/potem – zdaj/zdajle*, *noter/notri – ven, tukaj/tule – tam, gor – dol, sem* – (*tja* – 122. mesto na seznamu ključnih lem), *nazaj* – (*naprej* – 111. mesto). Da so v govorjeni rabi bolj pogoste kot v pisni, lahko pojasnimo z večjo vpetostjo govorjene rabe v prostor in čas kot komunikacijski okoliščini (kot tudi nasploh z večjo vpetostjo govorjene rabe v kontekst).

Tesno povezani z izrazi za orientacijo v prostoru in času so kazalni zaimki. Na seznamu prvih 100 ključnih lem so, poleg že prej naštetih *tukaj, tule* in *tam*, še: *tale, tak, takole, tako, tisti*. Razlaganje njihove uvrstitve med 100 za govorjeno rabo najbolj ključnih lem je težavnejše, saj so rabe precej raznovrstne. Vseeno lahko ugibamo, da podobno kot za izraze, s katerimi se orientiramo v prostoru in času, na to veliko vpliva njihova deiktična funkcija.

Prav tako sta na seznamu prvih 100 ključnih lem še osebna zaimka *jaz* in *oni*, kjer podobno pogosto izstopa deiktična funkcija. Vsaj pri zaimku jaz vpliva na frekventnost rabe tudi odsvetovano eksplicitno izpostavljanje prve osebe v pisni normi (tip *sej jz sn odgovarjal* proti *saj sem odgovarjal*) – s tega vidika je povezan s skupino slogovno zaznamovane leksike.

4.5 Izrazi za govorna dejanja, pogosta v govorjeni rabi

Poleg številnih kazalnih in dveh osebnih zaimkov se na seznamu 100 najbolj ključnih lem za govorjeno rabo znajde še kar nekaj vprašalnih zaimkov: *kaj, koliko, kje, kakšen, kako, kdaj*. Predpostavljamo, da je glavni razlog njihova raba v vprašanjih – to govorno dejanje je v interaktivni govorni komunikaciji pogostejše kot v pisni, ki je prevladujoče neinteraktivna.

Ostale leme, povezane z govornimi dejanji, bolj pogostimi v govorjeni kot pisni rabi, so še: *jutro* (*dobro jutro* – *pozdravljanje*), hvaliti (*hvala* – *zahvaljevanje*), *gospod* (nazivanje), *prositi* (*prosim* – vljudnostne konvencije).

Izstopa tudi skupina glagolov: to so glagoli premikanja (*iti, priti, hoditi*), rekanja (*reči, praviti*), dajanja (*dati*), delanja (*narediti, delati*), poslušanja (*čuti; slišati* na 135. mestu), gledanja (*videti, gledati, pogledati*), vedenja (*vedeti, znati*) in mišljenja (*misliti*) ter glagol *imeti*. Nekateri od teh (zlasti *vedeti, misliti, gledati, pogledati, videti*) dosegajo visoko frekventnost tudi ali celo zlasti zaradi diskurzno-označevalnih vlog, za ostale pa je verjetno razlog, da so aktivnosti oz. stanja, ki jih ti glagoli opisujejo, v govorjeni rabi pogosteje predmet zanimanja udeležencev komunikacije.

Večja skupina lem je takih, ki jih lahko pogosto povežemo z izražanjem odnosa, mnenja, stališča ali emocij govorca: *ful, fajn, super, res, joj, pizda, lepo, dobro, jah*. Tudi za izražanje tega predpostavljamo, da v govorni interakciji pogosteje najde prostor kot v pisni rabi.

4.6 Izrazi, »priljubljeni« v govorjeni rabi

Naslednja večja skupina lem je večinoma nepolnopomenska leksika, za katero se zdi, da je – morda tudi zaradi nekoliko specifičnih načinov rab in kolociranja – v govorjeni rabi bolj priljubljena in rabljena pogosteje, v dodatnih, novih pomenih in funkcijah. To so: *pač, malo, čisto, itak, nekako, točno, dosti, res, mogoče, zraven, lepo, sigurno, dejansko, stvar*, v določenih rabah tudi zaimek *kakšen* ter nedoločni zaimki *nekako, nek* in *nekje*. Med razlogi za njihovo večjo frekventnost v govorjeni rabi lahko prepoznamo tudi naslednje:

- pokažejo se nekatere dodatne funkcije oz. se rabe širijo na pomene, kjer bi v skladu z jezikovno normo uporabili kakšne druge sopomenske izraze, in tudi v resnici opazimo v pisnem korpusu obratno razmerje v frekventnosti, preračunano na 1 mio. pojavnic (če ilustriramo: v govornem korpusu je *sigurno* rabljeno pogosteje kot *zagotovo*, v pisnem korpusu je

zagotovo rabljeno pogosteje kot *sigurno*), npr. *malo* vs. *nekoliko*, *na kratko* (*najprej bi malo ponovili drugo svetovno vojno*); *čisto* vs. *povsem, popolnoma, zelo* (*jz sem čis lepo povedal kako in kaj*); *dosti* vs. *precej, dokaj* (*ena taka dost nevsiljiva regulacija*); *mogoče* vs. *morda* (v govornem korpusu je frekventnejši *mogoče*, v pisnem *morda*); *zraven* vs. *poleg* (v govornem korpusu je frekventnejši *zraven*, v pisnem *poleg*); *sigurno* vs. *zagotovo*; *dejansko* vs. v *resnici;*

- mnogi imajo ob tem neke vrste funkcijo relativiziranja, za katero se zdi, da je v govorjeni rabi bolj prisotna kot v pisni (gl. npr. Chafe in Danielewich 1987): *pač, malo, nekako, dosti, nek, mogoče, dejansko, nekje* (*včasih je blo mal težko pridet do avtomobila; tko da to so nekak dve bistveni spremembi; ja to je verjetn dost tak tip ne; edin kar nam manjka je v bistvu mogoč mal elegance; ne bo nekih večjih problemov; date v mikrovalovko za kakšno minuto; hitro po prvem marcu takrat nekje drugega marca*);
- nekaterim lahko pripišemo še kakšne druge diskurzne funkcije, npr. poudarjalno, interakcijsko ipd.: *itak, točno, res, lepo, sigurno* (*itak je vse predrago; res tis k da nam on sporoča to na ta način; tako ti dej lepo tele eee špagete kuhat; aja točno ja*);
- ponekod zasledimo tudi posamezne za govorjeno rabo tipične ali specifične kolokacije ali (pragmatične) frazeme: *čakaj malo, a res, lepo pozdravljeni, lepo prosim, lepo se imej/te, lepo te/vas prosim, ena stvar, prva stvar, druga stvar, kakšna stvar, taka stvar, pa te stvari …*

V tej skupini za govorjeno rabo ključnih lem torej opazimo tudi nekatere vrste rab, ki so za pisno rabo manj značilne ali celo slogovno ali normativno neprimerne, so pa značilne za govorjeno rabo.

4.7 Oblike, bolj »priljubljene« v govorjeni rabi

Na seznamu ključnih lem korpusa Gos opazimo, da se večkrat pojavljajo zaimki na *-le*. Med prvimi 100 ključnimi lemami so to *tule, tale, zdajle, takole*, v nadaljevanju seznama do 300 se zvrstijo še trije: *tukajle, tamle, takle*, naprej pa še nekateri (*tistile, prejle* …). Zdi se, da je taka oblika zaimkov v govorjeni rabi bolj »priljubljena« kot v pisni.

Nižje na seznamu ključnih lem, od 100 naprej, se začnejo večkrat pojavljati tudi različne pomanjševalnice, npr. *pesmica, slikica, kavica, mamica, majčka, flaška, dudka, mucka, zadevica, otroček, pobič, gvantek, kuglica, dekica, šipica, mlekec, župica*. Podobno kot pri zaimkih na *-le* se vsiljuje hipoteza, da so pomanjševalnice v govorjeni rabi nekoliko bolj »priljubljene« kot v pisni.

4.8 Drugo

Na seznamu prvih 100 ključnih lem se pojavi tudi nekaj veznikov: *ampak, samo, ker, saj, torej*. Izražajo večinoma protivnost, vzročnost, posledičnost in sklepalnost. O razlogih za njihovo pogostejšo rabo v govorjeni rabi težko sklepamo, potrebna bi bila celostnejša primerjalna analiza rabe veznikov v govorjeni in pisni rabi.

Na seznamu prvih 100 ključnih sta tudi števnika *trideset* in *dvajset* ter v govornem korpusu zelo pogosto z njima povezan pridevnik *cel* (tip *ena cela sedem*). V nadaljevanju seznama se takoj po prvih 100 ključnih lemah pojavlja še več števnikov (*petdeset, sto, osemdeset, devetdeset* ...). Razlog je najverjetneje v tem, da so v govornem korpusu vsi števniki izpisani in nikoli zapisani s cifro.

5 DISKUSIJA

Potem ko smo poiskali osrednje in pogoste vzroke, zakaj so določene leme v govorjeni rabi frekventnejše kot v pisni, nas zanima, katere od teh lahko dejansko povežemo s samim govornim kanalom ter posledično leksikalne enote in/ali njihove posamezne rabe v slovarju označimo kot tipične za govorjeno rabo. V pregledu literature v razdelku 2 smo povzeli zaključek Akinnasa (1982), da nekatere komunikacijske potrebe večinoma zadovoljujemo skozi govorjeno rabo, druge prevladujoče skozi pisno rabo, in da te funkcijske razlike vplivajo na leksikalne, skladenjske in semantično-pragmatične lastnosti besedil v enem in drugem načinu jezikovne rabe. Ali torej lahko ločujemo med tem, kaj je tipično za govorjeno rabo zaradi lastnosti samega govornega kanala, in tem, kaj je na seznamu ključnih lem zaradi funkcijskih razlik med obema kanaloma? Odgovor na to vprašanje smo iskali tako, da smo skušali definirati lastnosti govornega kanala in jih ločiti od diskurznih funkcij, pogosto povezanih z govornim kanalom.

Verjetno najbolj prepoznavna fizična lastnost govornega kanala je izgovarjanje. Z njim so na leksikalni ravni povezana popravljanja, neverbalno odzivanje (*eee, mhm* ...) ter zapolnjevanje vrzeli, ki lahko vključuje tudi povsem leksikalizirane elemente (*v bistvu, da tako rečem*) in ni vedno nedvoumno prepoznavno.

S fizičnimi lastnostmi govornega kanala tesno povezana lastnost govorjene rabe je prevladujoča spontanost tvorjenja. Takoj ko nekaj izgovorimo, je to hkrati že preneseno naslovniku: vmes ni veliko časa za razmišljanje in popravljanje. Toda hkrati vemo, da se v javnem diskurzu besedilo, ki bo govorjeno, pogosto

pripravi vnaprej (zapiše), ali pa se vsaj izdelajo vsebinske oporne točke oz. lahko govorjena raba poteka ob hkratni pisni ali multimedijski, npr. predstavitve ob prosojnicah. Spontanost zato ni več tako vseprisotna lastnost govorjene rabe kot izgovarjanje.

Govorni kanal zahteva hkratno prisotnost tvorca in naslovnika, s tem pa omogoča sproten odziv naslovnika in neposredno interakcijo med njim in tvorcem. Lahko bi torej rekli, da je lastnost govornega kanala tudi interaktivnost. Toda v diskurzu prek javnih medijev to, razen izjemoma, ne velja, prav tako ne v drugih oblikah javnega govorjenega diskurza, kot so npr. gledališče, javne prireditve ipd. Vedno ko imamo na eni strani zelo množičnega naslovnika, je interaktivnost govorjene rabe precej ali celo popolnoma okrnjena.

Prav tako bi lahko trdili, da je govorjena raba bolj vpeta v prostor in čas in nasploh v kontekst ter da je tudi to lastnost, ki izhaja iz samega govornega kanala. Toda tudi pisna jezikovna raba ima svoje prostore in čase, v katere je vpeta, le da je ta lastnost nekoliko manj izrazita in eksplicitno prisotna.

Na skrajnem drugem koncu je vprašanje neformalnosti govorjene rabe. Trdili bi lahko, da izgovarjanje in poslušanje zahtevata bližino tvorca in naslovnika v prostor-času, zaradi česar govorjena raba praviloma poteka med manj udeleženci, torej je posledično pogosteje intimnejša in zato manj formalna. Toda mnogo situacij govorjene rabe v javnem diskurzu je prevladujoče formalnih, zaradi družbenih razmerij pa so take tudi nekatere nejavne nezasebne situacije, npr. razgovori za službo, konzultacije pri profesorju, posredovanje informacij, tudi delovni sestanki itd.

Po tem razmisleku lahko sklenemo, da ne moremo enostavno ločevati lastnosti govorjene rabe od diskurznih funkcij, ki jih skozi to rabo bolj pogosto ali celo praviloma uresničujemo.

Navedene lastnosti govornega kanala skušamo v Tabeli 1 povezati s skupinami izrazov, kot smo jih opisali v razdelku 4. V stolpcu »skupina« navajamo skupine izrazov, definirane v razdelku 4, v stolpcu »funkcije v govorjeni rabi« povzemamo najznačilnejše funkcije, ki jih ti izrazi opravljajo v govorjeni rabi, v stolpcu »lastnosti govorjene rabe« pa skušamo navesti tiste značilnosti govorjene rabe, ki spodbujajo njihovo večjo frekventnost, pri čemer na prvem mestu navedemo najvplivnejšo lastnost.

Tabela 1: Tipična leksika govorjene rabe, njene funkcije in z njimi povezane lastnosti govornega kanala.

Skupina	Funkcije v govorjeni rabi	Lastnosti govorjene rabe
Neverbalni izrazi	neverbalno izražanje	izgovarjanje spontanost interakcija
Diskurzni označevalci	različne metadiskurzne vloge	interaktivnost spontanost
Slogovno zaznamovana leksika	nestandardni jezik	neformalnost
Deiktiki	deiktična funkcija	vpetost v kontekst
Izrazi za govorna dejanja, pogosta v govorjeni rabi:		
• vprašalni zaimki	spraševanje	interaktivnost
• glagoli	diskurzni označevalci izražanje predmetnosti	interaktivnost vpetost v kontekst
• izrazi emocij, drže ...	izražanje emocij, drže ...	neformalnost
Izrazi, »priljubljeni« v govorjeni rabi	novi pomeni in funkcije	neformalnost spontanost interaktivnost
Oblike, bolj »priljubljene« v govorjeni rabi	?	?

Z najbolj prepoznavno fizično lastnostjo govornega kanala, izgovarjanjem, so povezani samo neverbalni izrazi. Diskurzne označevalce in izraze za izražanje vprašanj smo povezali predvsem z interaktivnostjo, ki pa ni značilna za čisto vse oblike govorjene rabe in tudi ni ekskluzivna za govorjeno rabo. Slogovno zaznamovano leksiko, izraze emocij oz. drže ter izraze, »priljubljene« v govorjeni rabi, smo povezali predvsem z neformalnostjo. Tudi tukaj ne gre za lastnost, ki bi bila značilna za čisto vse oblike jezikovne rabe in ki bi bila lastna izključno govorjeni rabi. Deiktike in glagole za izražanje predmetnosti smo povezali z večjo vpetostjo govorjene rabe v kontekst. Za oblike, »priljubljene« v govorjeni rabi, nismo nobeni od zgoraj definiranih lastnosti govornega kanala pripisali, da bi posebej spodbujala njihovo večjo frekventnost.

Na podlagi tega razmisleka in poskusne shematizacije v Tabeli 1 je naš sklep, da ne moremo enostavno ločevati, katere ključne leme ali njihove posamezne rabe naj v slovarskem opisu označimo kot tipične za govorjeno rabo in katerih ne. Naš zaključek je, da je označevanje govorjenosti proti pisnosti preveč splošen kriterij, da bi bil primeren kot slovarski kvalifikator, razlog pa je ta, da sta tako pisna kot govorjena

raba zelo raznovrstni. Predlagamo, da se ustreznejše slogovno-zvrstne oznake leksikalnih enot in rab iščejo na bolj specifičnih ravneh, na primer: 'neformalno', 'v pogovoru', 'regionalno omejeno', 'za izražanje emocij/drže' in podobno.

6 ZAKLJUČEK

V prispevku smo razpravljali o vprašanjih, katera leksika je tipična za govorjeno rabo in ali lahko tem leksemom ali posameznim načinom njihove rabe pripišemo kvalifikatorsko oznako »v govoru«.

Po pregledu tujih raziskav o razlikah med pisno in govorjeno rabo smo zavzeli stališče, da so na ravni leksike med obema kanaloma razlike že predhodno dovolj obsežno dokazane, da lahko brez zadržka predpostavimo, da govorni korpus nosi dodatne podatke, zlasti na ravni leksikalnih vzorcev in kolokacij, ki jih iz še tako velikih pisnih korpusov ne moremo pridobiti.

Nato smo predstavili, katere skupine leksemov so v korpusu Gos tipično pogosteje rabljene kot v referenčnem korpusu Gigafida, v diskusiji pa je bila v središču našega zanimanja dilema, ali lahko tem rabam pripišemo kvalifikatorsko oznako govorjenosti. Dilema izhaja iz dejstva, da je govorjena raba pogosto povezana z drugimi funkcijami in situacijami kot pisna ter da razlike med eno in drugo rabo niso nujno pogojene z govorjenostjo proti pisnosti, ampak z drugimi diskurznimi faktorji. Ko smo skušali definirati lastnosti govorjene rabe, ki izhajajo iz samega govornega kanala, le-teh nismo mogli jasno ločevati od višjih diskurznih lastnosti, npr. neformalnosti, interaktivnosti itd., predvsem pa razen lastnosti izgovarjanja nobene druge nismo mogli pripisati celotnemu spektru različnih tipov govorjene rabe niti jih nismo mogli prepoznati kot imanentnih izključno govorjeni rabi. Zato smo zaključili, da govorjenost proti pisnosti ni ustrezna kategorija za označevanje leksikalnih enot, ker je presplošna, in predlagali iskanje ustreznih oznak na drugih, ožje opredeljenih ravneh, kot so 'neformalno', 'v pogovoru', 'regionalno omejeno', 'za izražanje emocij/drže' in podobno.

VII
Specializirana leksika in splošni slovar

Specializirana leksika v splošnem slovarju

Špela Vintar

Abstract
This paper discusses theoretical and methodological issues related to specialised vocabulary in a dictionary of contemporary Slovenian. Key issues are addressed such as the role of terminology in a general dictionary, user requirements and needs, the complexity of the distinctions between general and specialised terms, as well as corpus composition and representativeness. We propose a model where lexical items are categorised into three levels of termhood, with each level of specialisation requiring a different strategy of lexicographical description. By illustrating possible relations between the proposed categories and the corpus-based methodology of candidate extraction, a working methodology is established for handling specialised units in a general dictionary.

Keywords: specialised vocabulary, general dictionary, terminology extraction, user requirements

Ključne besede: specializirana leksika, splošni slovar, luščenje terminologije, uporabniške zahteve

1 UVOD

V vsakdanjem življenju vsi uporabljamo specializirane izraze, saj se prav vsak od nas ukvarja ali srečuje s kako dejavnostjo ali življenjskim področjem, ki ga ne delijo vsi govorci slovenščine in ki zahteva določena znanja ali izkušnje – je torej specializirano. Ob tem smo kot materni govorci opremljeni z intuitivno zaznavo, da so nekateri izrazi bolj specializirani od drugih, svoj senzor terminološkosti pa znamo nasloniti tudi na konkretne in uporabne utemeljitve: »Te besede nihče ne razume« ali »Tako temu rečejo mehaniki«.

Prispevek predstavlja niz razmišljanj o vključevanju specializirane leksike v slovar sodobnega slovenskega jezika, vanje pa so zajeti različni vidiki od leksikografske tradicije pri nas in drugod, predvidevanj želja in potreb uporabnikov, zbiranju in izbiranju korpusnega gradiva pa vse do leksikografskega oziroma terminografskega opisa in predstavitve podatkov po meri uporabnikov.

Že na prvi pogled je jasno, da vzpostavitev celovitega metodološkega okvira za vključevanje specializirane leksike v slovar ni enostavna naloga, vendar bi pričakovali, da so na podobno temeljna vprašanja v preteklosti odgovarjale že leksikografske ekipe pomembnih splošnih enojezičnih slovarjev in da je njihova spoznanja razmeroma preprosto prenesti v naše okolje. A ko začnemo podrobneje pregledovati strokovno literaturo, je podrobnih in eksplicitnih metodoloških študij o specializirani leksiki v splošnih slovarjih presenetljivo malo, še posebej v primerjavi z bogatim naborom literature o metodologiji specializiranih in terminoloških slovarjev. O slovarskih konceptih in odločitvah v zvezi s specializirano leksiko nam tako še največ povedo sami slovarji in ponekod njihovi uvodi. Druga izrazitejša težava pri prenosu tujih izkušenj v slovensko okolje pa je specifičnost slovenske leksikografske tradicije na eni strani in aktualni družbeni ter jezikovnopolitični okvir na drugi strani, oboje namreč vpliva na pričakovanja uporabnikov slovarja in posledično na nabor funkcij, ki naj bi jih bodoči slovar opravljal. Iz tega izhaja osvobajajoče in hkrati zavezujoče prepričanje, da je metodologijo obravnave specializirane leksike v sodobnem slovarju slovenščine treba vzpostaviti na novo, z njenim preskušanjem v praksi pa naj bi se samodejno razvil tudi iterativni cikel metodoloških popravkov in izboljšav.

2 VLOGA SPECIALIZIRANE LEKSIKE V SPLOŠNEM SLOVARJU

Kot ugotavlja Boulanger (1996: 141) v svojem diahronem pregledu, je že od 19. stoletja naprej v leksikografiji opazna težnja po vse večjem vključevanju

strokovnega izrazja v splošne eno- in večjezične slovarje. Od razsvetljenstva naprej je vztrajno rasla vloga znanosti in tehnike v vsakdanjem življenju, prav tako se je zviševala raven izobrazbe in posledično vključenost t. i. tehnolektov v vernakularni jezik in jezikovne priročnike. Druga polovica 19. stoletja je bila tudi za slovenščino obdobje živahnega ustvarjanja strokovnega izrazja, pomembno zaznamovano tudi s prevodi znanstvenih in strokovnih del iz pretežno nemščine, pa tudi drugih jezikov (Prunč 2009).

V 20. stoletju so se slovarji začeli odmikati od normativne vloge in se približevati uporabnikom, ti pa so od slovarjev velikega obsega pričakovali vse boljšo zastopanost specializirane leksike. Landau (2001) ugotavlja, da so veliki splošni slovarji vse bolj podobni združevanju številnih specializiranih (LSP) slovarjev z nekdanjim splošnim, vzrok za to pa je po njegovem tudi dejstvo, da v segmentu specializirane leksike zaradi naglega razvoja znanosti in tehnologije nastaja bistveno več novega izrazja kot v splošnem besedišču. Utemeljitev za vse večje vključevanje specializirane leksike v slovarje je več, če jih povzamemo po Josselin-Leray (2005):

a) Leksikografska tradicija, ki – kot smo že omenili – vsaj zadnje stoletje in pol narekuje vse višji delež specializirane leksike v splošnih slovarjih (prim. razdelek 2.2).

b) Vse intenzivnejše vstopanje specializirane leksike v vsakdanji diskurz, ki ga imenujemo tudi determinologizacija ali celo vulgarizacija terminologije. Pri laični rabi terminologije pogosto prihaja do odstopanj od prvotnega specializiranega pomena, pa tudi metaforizacije (npr. *anoreksija obrestnih mer*); za splošni slovar je pomembno, da opisuje tudi takšno rabo.

c) Didaktična oz. pedagoška vloga splošnih slovarjev, ki je najprej temeljito preobrazila angleško slovaropisje predvsem zaradi potreb učencev angleščine kot tujega jezika, a v sodobnem globaliziranem svetu ta vloga vsaj deloma pritiče večini enojezičnih slovarjev.

č) Težnja po izčrpnosti (angl. *comprehensiveness*), ki želi z ustvarjanjem enega vira za različne namene zadostiti čimveč uporabniškim potrebam, obenem pa je ob sodobnih informacijskih tehnologijah in primarnosti elektronske oblike izčrpnost bistveno lažje dosegati kot nekoč.

d) Pričakovanja in potrebe uporabnikov, ti se namreč za tehnologijo in znanost vse bolj zanimajo in so vse bolje obveščeni.

Vsi navedeni razlogi obstajajo tudi v slovenskem prostoru. Če se pri leksikografski tradiciji omejimo le na SSKJ in njegovega načrtovanega naslednika NSSKJ, je SSKJ terminologiji namenil pomembno vlogo, pri tem pa ustrezno zamejil raven specializiranosti in aktualnost:

> Terminologija je upoštevana nekako v obsegu srednje šole, zlasti če jo podpira publicistična ali poljudnoznanstvena raba. /.../ Terminološko gradivo je nastalo deloma z izpisovanjem poljudnoznanstvenih del, srednješolskih učbenikov in strokovnih slovarjev, deloma pa s prispevki okoli sto terminoloških svetovalcev. Od tega gradiva so bili za slovar odbrani samo termini, ki se rabijo v novejšem času. (SSKJ, Uvod: XVI–XVIII).

Zastopanost posameznih ved v SSKJ ni uravnotežena, prav tako pogostnost rabe ni bila poglavitni kriterij za vključevanje terminologije v slovar, a z vključevanjem terminoloških svetovalcev so skušali leksikografi doseči najboljši možni kompromis med danimi omejitvami glede obsega in specializiranosti ter strokovnostjo razlag:

> V terminologiji so najbolj zastopane tehnične vede, manj pa obrti in posamezne stroke, za katere ni bilo mogoče dobiti popolnejšega gradiva (montanistika, radio in televizija, bibliotekarstvo), ali stroke, ki so pri nas šele v razvoju (pomorstvo, kibernetika). (ibid. XVI).

Metodološka izhodišča glede vključevanja terminologije v NSSKJ se v veliki meri naslanjajo na stari SSKJ, izpostavljena pa je potreba po drugačni zastopanosti področij, vključevanju novih področij in uporabi sodobnejših metod terminološkega posvetovanja (Humar 2009). V konceptu NSSKJ (Gliha Komac et al. 2015: 49–51), objavljenem marca 2015, je leksikografska metodologija nadrobneje obrazložena, in sicer avtorji specializirano leksiko delijo na povsem in delno determinologizirano, pri čemer je prva deležna običajne leksikografske obravnave brez dodajanja terminoloških kvalifikatorjev in sodelovanja področnih strokovnjakov, druga skupina pa predvideva oblikovanje posplošenih, a strokovno pravilnih razlag ob pomoči področnih strokovnjakov ter označevanje specializiranosti s terminološkimi kvalifikatorji. Kriterija za razločevanje med prvo in drugo skupino sta dva, in sicer stopnja determinologizacije ter poznanost izraza splošnemu uporabniku, pri ugotavljanju obeh pa naj bi se leksikograf naslanjal na podatke iz korpusa. Pri tem, po naši oceni ključnem metodološkem vprašanju, je omenjeni osnutek zelo nedorečen.

Če se vrnemo k razlogom za vključevanje terminologije v splošne slovarje, je postopek determinologizacije, ko v splošno rabo prehaja velik del inventarja posameznih strok (informacijske tehnologije) ali pa postajajo strokovna področja predmet širšega zanimanja (zdravje, finance, okolje, šport), intenziven tudi v sodobni slovenščini. Značilnost determinologizacije je, da izrazi močno razširijo svoje prvotno pomensko in ekspresivno polje, ob tem pa ohranijo zgolj del strokovnega pomena. Temu ustrezen mora biti slovarski opis, saj lahko predvidevamo, da bo slovenski uporabnik sodobnega slovarja z njim želel zadostiti tudi informacijskim potrebam po razlagah specializiranega besedišča, vendar pri pomenih, ki so prešli

v splošno rabo, ne bo pričakoval formalne terminološke definicije. Pri ugotavljanju stopnje determinologizacije nam predvsem pomaga korpusna metodologija, ki meri razširjenost izraza v povsem splošnih, poljudnostrokovnih in strokovnih besedilih, pomemben podatek pa je tudi pojavljanje izraza v uporabniško ustvarjenih spletnih besedilih ter v govornem korpusu.

Novi slovar sodobne slovenščine bo imel zagotovo tudi didaktično vlogo, pravzaprav bi močno potrebovali slovar, ki bi se tej vlogi prednostno posvetil. Vprašanje, kako naj slovar najbolje poskrbi za potrebe učečih se, najsi bodo rojeni govorci slovenščine ali tujci, presega okvir tega prispevka, se pa uporabniškim potrebam podrobneje posvetimo v naslednjem razdelku.

2.1 Potrebe in pričakovanja uporabnikov

Sodobni splošni slovar je zamišljen primarno v elektronski obliki in nagovarja več tipov uporabnikov, ki jih Arhar Holdt razvršča v tri skupine glede na namen uporabe: a) izobraževanje, kamor sodijo uporabniki slovenščine kot prvega, drugega in tujega jezika, b) z jezikom povezani poklici, kamor se uvrščajo lektorji, novinarji, prevajalci, književniki, akademiki, znanstveniki itd., in c) prosti čas, kjer najdemo jezikovne navdušence, ugankarje in reševalce križank, ljubiteljske jezikovne svetovalce in podobno (Arhar Holdt 2015).

Čeprav se z vprašanjem uporabniških tipov in njihovih slovarskih potreb v digitalni dobi ukvarjajo številni avtorji (Müller-Spitzer 2014; Hult 2014; Lew et al. 2014), je razmeroma malo teh raziskav usmerjenih prav v specializirano leksiko. Ob analizi ugotovitev različnih uporabniških študij in anket pa se izkaže, da so potrebe in pričakovanja uporabnikov v večini slovarskih situacij neločljivo povezani s specializirano leksiko; tako denimo Müller-Spitzer (2014) identificira pomembno, a pogosto spregledano uporabniško skupino jezikovnih entuziastov, ki se ljubiteljsko ukvarjajo z nenavadnimi besedami, rešujejo in sestavljajo križanke ali igrajo Scrabble, pri tem pa v slovarju pogosto iščejo prav specializirane izraze ali tujke. Hult (2014), ki je analizirala komentarje uporabnikov ob spletni različici pedagoškega slovarja švedščine, navaja številne kritike, ki se nanašajo na slabo zastopanost specializiranih izrazov v slovarju.

Posebej podrobno se s pričakovanji uporabnikov glede specializirane leksike ukvarjata Josselin-Leray in Roberts (2005), ki poročata o rezultatih ankete o uporabi eno- in dvojezičnih slovarjev, izvedene med angleško in francosko govorečimi uporabniki. Na vprašanja je odgovorilo skupaj 649 oseb iz treh uporabniških skupin: prevajalci, strokovnjaki in splošna javnost, ki v prispevku žal ni podrobneje opredeljena. Vprašanja so se nanašala na različne vidike uporabe slovarjev, skoraj

vsa – skupaj jih je bilo 19 – pa so bila povezana z deležem specializirane leksike. Če se omejimo le na enega od številnih vidikov, ki jih obravnavata avtorici, in sicer na vprašanje, ali se zdi uporabnikom delež vsebovane terminologije pomemben kriterij pri nakupu splošnega slovarja in ali bi kupili slovar brez terminov, so bili rezultati enoznačno v prid terminologije, in sicer pri vseh skupinah uporabnikov, tudi pri laični, saj je v povprečju kar 60 % vprašanih kategorično zavrnilo možnost nakupa slovarja brez specializirane leksike.

V Sloveniji sorodne raziskave, ki bi se ciljno ukvarjala s pričakovanji in potrebami uporabnikov slovarja glede terminologije, še ni bilo, zato skušamo to vrzel premostiti z analizo strežniških dnevnikov, pridobljenih z nekaterih terminoloških spletišč (Vintar 2015a).

2.2 Delež specializirane leksike

Če predpostavimo, da si torej tudi slovenski uporabnik v slovarju želi specializirano leksiko, je naslednje naravno vprašanje, kolikšen delež besedišča naj bi predstavljale specializirane iztočnice v razmerju do splošnega jezika in do celotnega obsega slovarja.

Pregled tuje literature tudi na to vprašanje ne daje enoznačnih odgovorov. Landau (1974) pri analizi Webstrovega slovarja Third New International Dictionary omenja približni delež 40 %, vendar analiza posameznih strani slovarja prikaže vse do 89 % specializiranih gesel. Béjoint se prav tako ukvarja s tem vprašanjem, a se o deležu v odstotkih ne opredeli zaradi zahtevnega določanja meje med termini in netermini (Béjoint 1988: 360). S podobno težavo sta se spopadla tudi Boulanger in L'Homme (1991: 25); v kasnejši študiji pa Boulanger ugotavlja za angleške in francoske enojezične slovarje med 40 in 50 % specializirane leksike (Boulanger 1996: 147). Novejša študija Vrbinc in Vrbinc (2013) raziskuje zastopanost specializirane leksike ter uporabo področnih oznak v slovarjih Oxford Advanced Learner's Dictionary tretje, četrte in osme izdaje. Avtorici ugotavljata postopno zviševanje deleža specializiranih leksemov, a ne omenjata številk, ugotavljata pa tudi posodobitve in izboljšave v rabi področnih oznak, ki jih je sicer v OALD8 devetnajst.

Tu zastopani pristop h gradnji sodobnega slovarja skuša biti pragmatičen v smislu, da mora biti izbrani metodološki okvir glede zajema in obdelave specializirane leksike časovno in finančno izvedljiv, obenem pa učinkovit in trajnosten. Digitalna oblika omogoča popolno svobodo glede količine ponujenih podatkov, računalniške metode luščenja terminologije in definicij iz korpusov pa močno olajšajo pot do leksikografskih podatkov. Ker specializirani izrazi, dokler so pasivno shranjeni

v slovarski bazi in tam čakajo na ustrezno poizvedbo, nikogar ne motijo, bi bili teoretično glede njihovega deleža lahko nadvse radodarni. Še vedno pa je leksikografska obdelava iztočnic zahtevno strokovno delo, ki pri specializiranih enotah zahteva tudi sodelovanje s strokovnjaki, zato so omejitve v luči izvedljivosti slovarskega projekta še kako nujne. Poleg tega ni točno, da velika količina redundantnih specializiranih podatkov v slovarski bazi uporabnika ne moti, saj pri vsaki odprti poizvedbi te iztočnice povzročajo šum in dajejo popačeno sliko tega, kar naj bi slovar splošnega jezika predstavljal.

Ker si je v kontekstu korpusne leksikografije razmeroma nesmiselno vnaprej določiti, kolikšen delež slovarskih iztočnic naj bi bil specializiran, saj bi bilo to podobno, kot če bi pri jemanju vzorca krvi zdravnik vnaprej določil število rdečih in belih krvnih telesc, nam na zgoraj zastavljeno vprašanje ostane le zelo krhek odgovor: splošni slovar naj vsebuje ravno pravšnji delež specializiranih enot, da bo z njim zadovoljil kar največ informacijskih potreb, obenem pa ohranil učinkovitost poizvedovanja in karakter splošnega slovarja.

Naslednje vprašanje, ob katerem se bežno ustavimo, je, ali naj bodo posamezna specializirana področja v slovarju zastopana v enakih oziroma uravnoteženih deležih. Leksikografska tradicija kaže, da si tega cilja veliki splošni slovarji navadno niso zastavljali, večinoma so se mu celo izrecno odpovedali iz različnih razlogov (Josselin-Leray 2005: 146). Nemalokrat so na dobro ali slabo zastopanost posameznih strok v slovarjih vplivali tudi povsem idiosinkratični dejavniki: »Eden od urednikov Oxfordovega slovarja angleščine je bil slučajno amaterski mineralog in zato je akademski Oxfordov slovar angleščine še posebej bogat z mineraloško terminologijo« (Béjoint 1988: 361; prevod Š. V.). Če bi želeli zadostiti pogoju uravnoteženosti, bi se namreč takoj soočili z dejstvom, da stroke med seboj niti približno niso primerljive glede količine specializiranih izrazov, ki jih uporabljajo, zato bi bila kakršna koli številčna uravnoteženost nesmiselna. Poleg tega so nekatere stroke za medije in širšo javnost bolj zanimive kot druge, posledično iz takšnih strok več izrazja prehaja v neterminološko rabo (in v splošne korpuse). Ahmad et al. (1995) menijo, naj bi se splošni slovar osredotočal na tiste specializirane izraze, pri katerih se laični in strokovni interes prekrivata. S tega zornega kota bi v splošnem slovarju po vsej verjetnosti pričakovali izraze, kot so *dokapitalizacija*, *hipertenzija* in *prisilna poravnava*, manj pa *magnon* (fizika), *manipul* (zgodovina) ali *izobutil acetat* (kemija).

Privzemimo torej, da splošnemu slovarju ni treba uravnoteženo predstavljati vseh strok tega sveta, namesto tega pa naj bi zajemal izrazje vseh tistih strok, s katerimi se bolj ali manj pogosto srečuje tudi laik. Ob tem privzamemo tudi, da splošni slovar za nobeno od teh strok ne more dosegati enake ravni poglobljenosti in celovitosti kot terminološki slovar (Peters in Fernandez 2013). Če besedišče slovarja

zajemamo iz velikega reprezentativnega korpusa, kakršen je Gigafida, bi lahko domnevali, da je že pojavnost izraza v korpusu dokaz za »srečevanje laičnega in strokovnega interesa« in se tako pri presojanju o vključevanju specializiranih enot v slovar naslanjali kar nanjo. V naslednjem razdelku bomo skušali pojasniti, zakaj tako poenostavljen pristop ne zadošča.

3 GRADIVO ZA ZAJEM SPECIALIZIRANE LEKSIKE

V nadaljevanju v razmislek o metodologiji vključevanja specializirane leksike v splošni slovar uvedemo še najpomembnejši in hkrati najobčutljivejši parameter, to je korpus.

3.1 (Ne)uravnoteženost korpusa in (ne)uporabnost absolutne pogostosti

Že od začetkov korpusnega jezikoslovja smo priča tehnološkemu razvoju, ki omogoča gradnjo vse večjih korpusov, od razmaha interneta pa so digitalna besedila postala tako vseprisotna, da korpusi z milijardo ali več pojavnicami niso več nikakršna redkost.

Za gradnjo korpusnih splošnih slovarjev uporabljamo referenčne korpuse, ki naj bi – če so tudi dobro uravnoteženi – predstavljali nekakšen skupni imenovalec različnih zvrsti in podzvrsti, ki jih uporabljajo govorci nekega jezika. Za slovenščino imamo tako na voljo Gigafido s prek milijardo pojavnic, ki ni uravnotežena, ter Kres s 100 milijoni pojavnic, ki je vzorčni korpus z izenačenimi deli posameznih žanrov; oba skupaj tvorita tudi vir za ugotavljanje in primerjavo pogostosti leksikalnih enot, kar je za leksikografa pomembna informacija.

Pri zajemu specializirane leksike nam splošni referenčni korpus lahko nudi pomembne dokaze o tem, ali je določeni specializirani izraz prodrl tudi v polstrokovno oziroma povsem nespecializirano rabo, medtem ko so podatki o pogostosti uporabni zgolj z zadržkom.

	GF	Kres
paralelogram	107	18
trapez	923	126
deltoid	23	3

Matematični izrazi *paralelogram*, *trapez* in *deltoid* so istoredni termini, ki označujejo like v poglavju srednješolskega učbenika o večkotnikih, a se v korpusu pojavljajo z zelo različnimi frekvencami. Če bi želeli v slovarju pravično zajeti matematično izrazje na srednješolski ravni, bi morali vse tri izraze vključiti, čeprav se pojavljajo s tako različnimi frekvencami. Pri izbranem primeru se hitro pokaže temeljni izziv zajemanja in slovarske obdelave specializiranih iztočnic, *trapez* ima namreč toliko višjo frekvenco zaradi svojih nematematičnih pomenov. V Gigafidi najdemo še naslednje specializirane pomene: polje omejitve (košarka; »podaja iz trapeza«), telovadno orodje (gimnastika; »artistka na trapezu«), pripomoček za dvig z bolniške postelje (medicina; »vstati s pomočjo trapeza«), nastavitev digitalne slike na monitorju (računalništvo; »pogrešam digitalno korekcijo trapeza«), človek za protiutež pri nagibu jadrnice (jadranje; »flokist visi v trapezu«).

Po vsej verjetnosti se bo leksikograf pri obdelavi iztočnice *trapez* moral pri vsakem pomenu posebej odločati, ali ga zapisati v slovar ali ne. Tudi če bi se pri tem lahko opiral na samodejno označene korpusne pogostosti posameznih pomenov, česar jezikovnotehnološka infrastruktura trenutno še ne omogoča, ostaja vprašanje reprezentativnosti in uravnoteženosti korpusa za izbrane specialnoleksikografske namene odprto. V trenutni korpusni situaciji je namreč povsem mogoče, da se računalniški pomen pojavlja nesorazmerno pogosto zgolj zato, ker korpus vsebuje precejšnje število besedil iz revije Monitor, tam pa uredniki redno izvajajo testiranja monitorjev in poročajo o njihovi korekciji trapeza.

Če torej za nespecializirano leksiko lahko pričakujemo, da nam bodo podatki o pogostosti v uravnoteženem referenčnem korpusu predstavljali relevantno leksikografsko osnovo, tega ne moremo z gotovostjo pričakovati za specializirane izraze.

3.2 Prednostna področja in dopolnjevanje gradiva

Opisana situacija nas pripelje do sklepa, da če slovar ne namerava zajemati vseh področij in se po drugi strani pri zajemu specializiranih iztočnic ne more zanašati zgolj na korpusno pogostost, je treba v temeljnem slovarskem konceptu opredeliti tista strokovna področja (v povezavi z žanri in ravnijo specializiranosti), ki jih bo slovar zajemal. Za izbrana področja s povezanimi žanri in registri je treba zagotoviti dovoljšen obseg gradiva, da bo mogoče izvajati avtomatske postopke luščenja in jezikovnotehnološke obdelave z namenom pridobivanja iztočnic ter pripadajočih leksikografskih podatkov.

Pri odločanju, katera področja so prednostna, nas do neke mere vodijo temeljni cilji slovarskega projekta ali z drugimi besedami ciljne skupine uporabnikov, vendar je ob tem treba priznati, da bodo odločitve nujno intuitivne in subjektivne,

saj natančne in merljive opredelitve slovarskih potreb ciljne publike v našem prostoru še nimamo; pomemben prispevek k temu je sklop o uporabnikih v tej monografiji skupaj s prispevkom o analizi dnevnikov iskanja Termanie.

V procesu razmišljanja o prednostnih področjih so se izoblikovala tri načela, ki jim skušamo slediti. Prvo se navezuje na prej omenjeno leksikografsko načelo »stičišča med splošno in (zelo) specializirano rabo«, kar pomeni, da velja nameniti posebno pozornost tistim strokovnim področjem, ki so s svojim izrazjem močneje zastopana v splošnem javnem diskurzu in medijih. O njih nam zmore precej povedati že Gigafida, zato smo izvedli hiter pregled 1.000 najpogostejših samostalniških lem in jim pripisali področja. Nastal je naslednji grobi seznam področij: *politika, šport, pravo, ekonomija, finance, mediji, okolje, uprava, zdravje, kultura, IT, promet, turizem.*

Natančnejšo analizo izstopajočih tem v Gigafidi in primerjavo s korpusom slWaC sta izvedla Logar Berginc in Ljubešić (2013) z metodo tematskega modeliranja. V tej analizi se je osem tem, ki sta jih avtorja ročno poimenovala na podlagi samostalniških lem v vsaki temi, izkazalo kot skupnih obema korpusoma: notranja politika, finance, ekipni šport, vojna in terorizem, publikacije in kultura, lokalna (prostorska) politika, zdravje in pravo. Po drugi strani je nekaj tem opaznejših v enem ali drugem korpusu: v Gigafidi izstopajo naselje in cestni promet (zlasti prometne nesreče), prireditve, televizijski in radijski program, neekipni športi ter zaposlitev; v slWaCu pa film, potovanja in turizem, zunanja politika ter mali oglasi.

Z vidika metodologije moramo ob tem pripomniti dvoje. Lastni pregled tisoč najpogostejših lem iz Gigafide je bil opravljen z mislijo na specializirano leksiko, kar pomeni, da številnih lem nismo uvstili na nobeno tematsko področje, saj so se zdele splošne. Tematsko modeliranje Logar Bergінčeve in Ljubešića pa je bilo usmerjeno v prepoznavanje glavnih tematskih poudarkov in razhajanj med obema korpusoma ne glede na raven specializiranosti. Tema seveda ni enako strokovnemu področju, zato je pri morebitnih tovrstnih analizah z namenom dejanskega ugotavljanja zastopanosti posameznih strok v korpusu nujno najprej opredeliti metodologijo ter izbrati členitev področij. Drugi pomembni uvid pa se nanaša na dejstvo, da bo analiza zelo pogostih domnevno specializiranih samostalniških lem vključevala predvsem leksikalne enote, ki so delno ali povsem determinologizirane. V njih sicer prepoznamo izvorno stroko, a so splošno znane, za razumevanje ne terjajo poglobljenega poznavanja področja oziroma so postale pomensko ohlapnejše (*kredit, delnica, zaslon, liga, rak*).

Drugo načelo naslavlja bodoče slovarske uporabnike v izobraževanju in nadaljuje slovarsko tradicijo SSKJ s ciljem, naj novi slovar vključuje terminologijo vseh naravnoslovnih, družboslovnih in humanističnih strok na ravni

srednješolskih učbenikov. V tem segmentu obstoječa Gigafida ni reprezentativna, saj sicer vsebuje 75 osnovno- in srednješolskih učbenikov različnih založb in avtorjev, vendar so nekatera predmetna področja bistveno močneje zastopana kot druga. Za dobro pokrivanje srednješolske terminologije, ki bi jo v slovar zagotovo morali vključevati selektivno, je torej potrebna premišljena dopolnitev učbeniškega podkorpusa.

Tretje načelo pa izhaja iz spoznanja, da še tako obsežen in reprezentativen korpus ne more zajemati celotnega besedišča nekega jezika. Če pustimo ob strani govorjeno slovenščino, ki jo je v vsej bogati zvrstnosti tako rekoč nemogoče reprezentativno predstaviti v korpusu, so tu določene vrste pisne jezikovne rabe, ki so v obstoječih korpusih slabo zastopane. Identificirali smo eno takšnih vrzeli, ki smo jo označili s širokim izrazom življenjski dogodki, vanjo pa uvrščamo izrazje, povezano z različnimi pravno-upravnimi, socialnimi, verskimi in drugimi dogodki, s katerimi se vsaj občasno srečuje veliko število ljudi. Sem sodijo denimo različni bančni, zavarovalniški, poštni in upravni obrazci ter dokumenti, ki lahko vsebujejo pomembne specializirane izraze.

Pričakujemo tudi, da bi se ob podrobnejši korpusni analizi specializiranega besedišča posameznih strok pokazale nadaljnje vrzeli, saj uravnotežena zastopanost strok ter z njimi povezanih žanrov in registrov ni bila temeljna prioriteta korpusa Gigafida, poleg tega sestavo korpusa vselej določajo tudi nepredvidljivi vidiki pri pridobivanju besedil. Gigafida vsebuje 1.377 knjižnih stvarnih enot (4,24 %) in 5.871 enot iz revij (21,51 %), pri čemer niso vse revije specializirane. Sedanja besedilna taksonomija je preširoka, da bi omogočala vpogled v zastopanost posameznih strok, zato je v novi različici korpusa načrtovano podrobnejše tematsko označevanje.

Za tiste stroke, ki bodo za novi slovar prepoznane kot prioritetne, je v načrtu podrobni pregled obstoječega gradiva z navzkrižnim luščenjem izrazja ter dopolnjevanje s sodobnimi specializiranimi besedili. Po vsej verjetnosti bo treba vključiti še kakšno strokovno revijo ter poiskati ustrezne in sodobne strokovne publikacije, pri tem pa nimamo namena vključevati znanstvenih revij, monografij in disertacij. Za nekatere stroke bodo v pomoč tudi že obstoječi specializirani korpusi (turizem, odnosi z javnostmi; glej tudi http://nl.ijs.si/noske).

V sklopu tretjega načela, ki bi ga lahko imenovali tudi zavedanje o nepopolnosti korpusa, velja omeniti še neologizme. Novi izrazi nastajajo skoraj izključno v povezavi s specializiranimi področji, najmočneje opazimo poimenovalno potrebo pri novih tehnologijah, napravah in storitvah, ki sprožijo val zanimanja tudi med nespecializirano publiko in v splošnih medijih. Novim izrazom s korpusi težko sledimo, pa vendar je njihovo vključevanje v slovar posebej pomembno, še

posebej v primerih, ko nova poimenovanja ne predstavljajo naključnih avtorskih, uredniških ali prevajalskih domislic, temveč je pri njih prišlo do terminološkega dogovora ali vsaj refleksije.

Dober primer takšnega izraza je eden od razdelkov pričujoče knjige z naslovom *množičenje*. Angleški izraz *crowdsourcing* je bil najprej citatno prevzet in je v Gigafidi izpričan s štirimi pojavitvami, nato pa so se pričeli pojavljati poskusi slovenjenja: *moč množic, množicanje, množgančkanje, množičenje, množično zunanje izvajanje*. O najprimernejši ustreznici se je razvnelo kar nekaj (spletnih) razprav, mnenje je podala ekipa Terminologišča na Inštitutu Frana Ramovša za slovenski jezik ZRC SAZU, o izrazu so razpravljali uredniki in komentatorji slovarja informatike Islovar. Zaenkrat sicer ne moremo govoriti o dokončnem terminološkem dogovoru, čeprav sta oba prej omenjena foruma kot zmagovalca izbrala *množično zunanje izvajanje*. Nobena od možnih ustreznic v Gigafidi ni izpričana, pa vendar bi v novem slovarju pričakovali ustrezno leksikografsko obdelavo tega in sorodnega pojma *crowdfunding*.

4 ZAJEM SPECIALIZIRANEGA IZRAZJA

Če je še pred desetletjem ali dvema izdelava slovarskega geslovnika predstavljala najbolj garaško in dolgotrajno leksikografsko delo, nam danes jezikovnotehnološka orodja to delo močno olajšajo in pohitrijo, leksikografovi napori pa so bolj usmerjeni v pregledovanje, urejanje, čiščenje in dopolnjevanje samodejno izluščenih podatkov.

Za luščenje specializiranih enot iz besedil so se v preteklosti razvila številna orodja, tudi za slovenščino (Vintar 2010). Orodje LUIZ je bilo razvito za dvojezično luščenje izrazja iz angleško-slovenskih vzporednih in primerljivih besedil, lahko pa se uporablja tudi za enojezično luščenje. Osnovna zamisel luščenja temelji na primerjavi pogostosti med specializiranim in referenčnim korpusm in s tem sledi predpostavki, da so termini določenega področja pogostejši v besedilih tega področja kot v nespecializiranih besedilih. Tako ugotovljena »ključnost« za posamezne besede se kombinira z jezikovno odvisnimi oblikoskladenjskimi vzorci ter statistično hevristiko razvrščanja ter kot izhodne podatke ponudi seznam eno- in večbesednih terminoloških kandidatov.

Luščilnik LUIZ je bil v preteklosti preskušen za različna področja (Vintar in Erjavec 2008; Vintar in Fišer 2009; Vintar 2010; Logar et al. 2012; Pollak 2014), nikdar doslej pa z namenom pridobivanja izrazja za splošni slovar. Še preden smo se torej odločili luščilnik preskusiti za tukaj opisani namen, smo se dobro zavedali pomanjkljivosti pristopa, saj je zajem izrazja za specializirani slovar povsem

drugačna naloga od luščenja za splošni slovar. Eden od vidikov, ki je zahteval metodološki premislek, je že oblika terminov. Medtem ko v visoko specializiranem besedišču lahko pričakujemo veliko število tro-, štiri- in celo večbesednih terminoloških enot, je pri izrazju za splošni slovar primernejša omejitev na eno- in dvobesedne, izjemoma trobesedne enote. Prav tako je bilo treba temeljito prilagoditi hevristike glede pogostosti, saj redkih in visokospecializiranih izrazov ne želimo vključevati. O metodologiji polavtomatskega luščenja izrazja podrobneje razpravljamo v naslednjem razdelku.

S pomočjo luščenja ter navzkrižnih primerjav med podkorpusi resda pridobimo sezname potencialnih iztočnic po področjih, ki pa zahtevajo pregled in dopolnjevanje. Dopolnjevanje bo potrebno denimo pri novejših izrazih, ki bodo v korpusu izpričani s premajhno pogostostjo, da bi bili samodejno izluščeni, predvidevamo pa tudi uporabo sodobnih jezikovnotehnoloških metod za luščenje nad-, pod- in protipomenk ter zagotavljanje uravnotežene zastopanosti posameznih pojmovnih polj.

Že v tej fazi bo pri izdelavi področnih geslovnikov potrebna pomoč področnih strokovnjakov, ki bodo pri pregledu izluščenih seznamov prepoznali terminološke variacije, determinologizirane izraze in sopomenke ter s tem leksikografu olajšali razvrščanje izrazja v posamezne kategorije specializiranosti.

5 RAVEN SPECIALIZIRANOSTI IN LEKSIKOGRAFSKI OPIS

Že iz dosedaj opisanih primerov je razvidno, da niso vsi izrazi enako terminološki. Pri luščenju ključnih besed iz podkorpusa glasbenih učbenikov tako med prvimi 100 besedami opazimo mnogim znane glasbene izraze, kot so *ton, nota, sonata, mol, tempo, sopran*, pa tudi manj znane kompleksnejše izraze, pri katerih bi imeli kot laiki morda že težavo z oblikovanjem strokovno ustrezne razlage: *sinkopa, sekstakord, kromatičen, kadenca, modulacija*.

Če je izhodišče tu opisanega slovarskega projekta zadovoljiti najmanj tri heterogene skupine uporabnikov, tj. učence vseh stopenj, jezikovne profesionalce in tiste, ki jim je jezik konjiček, se morajo tudi tipi leksikografskega opisa prilagajati tako ciljni skupini kot specializiranosti izraza.

V skladu s tem smo predvideli razvrščanje specializiranih iztočnic v tri skupine, ki jim v nadaljevanju pravimo košarice in ki narekujejo različne pristope k leksikografskemu opisu.

Prva ali splošna košarica je najmanj specializirana in vključuje izraze, pri katerih sicer prepoznamo navezavo na določeno strokovno področje, vendar poznavanje področja ni pogoj za razumevanje in pravilno rabo. Ti izrazi so pogosti tudi v referenčnem korpusu (> 3.000) in se v splošnih besedilih rabijo povsem determinologizirano. Pri leksikografskem opisu ne potrebujejo področne oznake, niti se njihova specializiranost ne izraža v razlagi. Primeri takih iztočnic so *tempo, koncert, dirigent, nota, harmonija*.

Druga ali šolska košarica vsebuje izraze, ki jih redkeje srečamo v splošnih besedilih, a poimenujejo temeljne pojme stroke in se kot taki pojavljajo že v učbenikih nižje stopnje. Izkazujejo jasno navezavo na matično področje, a so izpričani tudi v referenčnem korpusu (> 300). Njihova razlaga lahko vsebuje navedbo področja, zlasti takrat, kadar se determinologizirani pomen razlikuje od področnega. Primeri takih izrazov so *sonata, akord, oboa, dur, mol, trozvok*.

Tretja ali strokovna košarica je najbolj specializirana, vsebuje izraze, ki so splošni publiki manj znani in je za njihovo razumevanje potrebno predznanje. Redko se rabijo v splošnih besedilih in se ne determinologizirajo; za slovar pa so vendarle pomembni z več vidikov: bodisi se pojavljajo v srednješolskih učbenikih in torej kljub specializiranosti sodijo na raven splošne izobrazbe, bodisi sodijo na eno od prioritetnih področij in je v referenčnem korpusu izpričana minimalna pojavnost. Pri njihovem leksikografskem opisu se uporabi področna oznaka, ki uporabnika opozarja na terminološkost, razlaga pa je strokovna. Takšni izrazi so denimo *septima, alikvotni ton, sekstakord*, če se omejimo zgolj na glasbo.

Če izraz ne izpolnjuje nobenega od navedenih kriterijev, se pravi ne izkazuje minimalne pojavnosti v splošnih besedilih, ne sodi na prioritetno področje niti med učbeniško izrazje, prav tako pa ne predstavlja nujnega dela pojmovnega sistema kakega sorodnega – pogostejšega – izraza, lahko sklepamo, da ni relevanten za splošni slovar.

Metodološko še najbolj zahtevni del opisane klasifikacije je učinkovita izraba korpusnih podatkov, saj se pri ugotavljanju specializiranosti določenega izraza nikakor ne moremo opirati zgolj na njegovo pogostost v referenčnem korpusu. Pri izrazih iz prve in druge košarice nam absolutna pogostost ne pove skoraj nič o dejanski pogostosti specializiranega ali iz njega izpeljanega determinologiziranega pomena, saj imajo številni izrazi več pomenov na različnih področjih. Tako je skupna pogostost izraza *sinkopa* v Gigafidi 269, vendar je od tega precejšnji del pojavitev v medicinskem pomenu sinkope kot začasne izgube zavesti ali omedlevice, drugi del se nanaša na za polovico podaljšani ton v glasbi, tretji del pa vsebuje še pojavitve v imenih podjetij, glasbenih skupin in preneseno rabo. Ustrezna leksikografska presoja se torej na pogostost lahko opira le po opravljenem postopku pomenskega razdvoumljanja.

6 ZAKLJUČKI

Dosedanji razmisleki predstavljajo nabor izhodišč, ki jih smiselno dopolnjujejo druga poglavja tega sklopa, obenem pa so izhodišča usklajena s temeljno filozofijo projekta novega slovarja sodobne slovenščine. Ob tem je – kot pri vsaki metodološki zasnovi – že vnaprej jasno, da se bodo sedaj grobo zarisani okvirji obravnave specializiranega besedišča zbrusili in konkretizirali šele skozi leksikografsko prakso, se pravi najprej v iterativnih ciklih samodejnega luščenja podatkov, nato pa v delovno intenzivnih postopkih urejanja, dopolnjevanja in obdelovanja slovarske vsebine.

Čeprav je vsak slovar že v izhodišču kompromis med idealnim in mogočim, pa si za novi sodobni slovar slovenščine želimo, da bi bilo glede specializirane leksike teh čim manj, kajti družba znanja, za katero si prizadevamo, se lahko uresničuje le skozi učinkovito ubesedovanje znanja, k temu pa lahko bistveno pripomorejo jezikovni priročniki.

Luščenje specializiranih izrazov za splošni slovar

Špela Vintar in Nataša Logar

Abstract
This paper describes an experiment aimed at extracting specialised vocabulary from specialised subcorpora for the purposes of general lexicography. The main objective was to test the methodology of automatic term extraction, which had been developed for specialised lexicography, in order to gain better insight into the weakness of, and the adjustments required to, the corpus in terms of domain representativeness. The LUIZ Term Extractor was used on two subcorpora, one containing a selection of texts pertaining to physics, biology and chemistry from the ccGigafida corpus, and the other a more homogeneous specialised corpus of textbooks on music theory. The results show that the domain of natural sciences, as represented in the Gigafida corpus, contains few lexical items which require special attention on account of their termhood, whereas a more specialised corpus, as expected, yields a larger number of highly specialised units.

Keywords: term extraction, specialised vocabulary, termhood, general language dictionary

Ključne besede: luščenje izrazja, specializirana leksika, terminološkost, splošni slovar

1 UVOD

Namen pilotnega eksperimenta, katerega rezultate prikazujemo v prispevku, je bil ugotoviti, ali so obstoječe metode luščenja specializiranega izrazja uporabne tudi za potrebe splošnega slovarja oz. kakšne prilagoditve zahtevajo. Dosedanje izkušnje z luščenjem (prim. razdelek 2.1) so bile namreč brez izjeme usmerjene v pridobivanje srednje- in visokospecializiranih izrazov za terminografske namene ali namene modeliranja strokovnih področij, tukajšnja naloga pa je precej drugačna, deloma celo povsem nasprotna.

Z eksperimentom smo želeli preliminarno odgovoriti na tri vprašanja, od tega sta bili dve povezani z gradivom:

1. Koliko gradiva s terminološkim potencialom pokaže Gigafida, če nanjo apliciramo metodo luščenja terminoloških kandidatov, ki je uspešna v specializiranih korpusih strokovnih besedil, na splošnih korpusih slovenščine pa še ni bila preizkušena?
2. Na kolikšen del Gigafide je metodo luščenja terminoloških kandidatov smiselno aplicirati?

Tretje vprašanje je bilo povezano s slovarjem, natančneje z uporabnostjo seznamov izluščenih terminoloških kandidatov pri razdelitvi terminološke (oz. neterminološke) leksike v tri košarice, ki smo jih predvideli za splošni slovar (prim. poglavje Specializirana leksika v splošnem slovarju):

Splošna košarica: Leksika, pri kateri sicer prepoznamo navezavo na določeno strokovno področje, vendar poznavanje področja ni pogoj za njeno razumevanje in pravilno rabo. Pri leksikografskem opisu ne potrebuje področne oznake, njena specializiranost se ne izraža niti v razlagi (*koncert, rastlina, ribič*).

Šolska košarica: Leksika, ki jo redkeje srečamo v splošnih besedilih, a poimenuje temeljne pojme strok in se kot taka pojavlja že v učbenikih nižje stopnje. Njena razlaga lahko vsebuje navedbo področja, zlasti takrat, kadar se determinologizirani pomen razlikuje od področnega (*lestvica, beljakovina, molekula*).[1]

Strokovna košarica: Leksika, ki je splošni javnosti manj znana in je za njeno razumevanje potrebno predznanje. Redko se rabi v splošnih besedilih in se ne determinologizira. Pri njenem leksikografskem opisu se uporabi področna oznaka, ki uporabnika opozarja na terminološkost, razlaga pa je strokovna (*alikvotni ton, hipertrofija, nazivna moč*).[2]

1 Ker so v šolski košarici srednjespecializirani izrazi, se predvideva, da v večini primerov razlago zanje lahko oblikuje leksikograf, pri čemer so v pomoč tudi učbeniška besedila, ki že sama vsebujejo veliko razlag in leksikografu služijo kot predloga.

2 S strokovno razlago mislimo na tip razlage, ki je oblikovno in konceptualno blizu terminološki definiciji in je strokovno pravilna, še vedno pa je prilagojena splošnemu uporabniku in je manj podrobna kot v terminološkem slovarju.

Tretje vprašanje se je torej glasilo: Kako lahko metoda luščenja terminoloških kandidatov leksikografom pomaga pri odločitvah o umestitvi leksike v posamezne kategorije specializiranosti splošnega slovarja?

2 METODA IN GRADIVO

2.1 Luščenje terminoloških kandidatov

Za luščenje smo uporabili luščilnik LUIZ (Vintar 2010), ki iz specializiranega korpusa izlušči terminološke kandidate na podlagi oblikoskladenjskih vzorcev, te pa nato razvrsti glede na t. i. terminološko utež. Slednja je hevristika, ki upošteva ključnost (Scott 1998) posameznih delov terminološke besedne zveze, njeno dolžino ter pogostost v specializiranem podkorpusu.

V dosedanjih raziskavah (Vintar in Erjavec 2008; Logar Berginc in Vintar 2008; Vintar in Fišer 2009; Logar et al. 2012) smo pri luščenju uporabljali obsežen seznam potencialnih oblikoskladenjskih vzorcev za slovenščino, ki se sicer lahko prilagaja specifikam posamezne stroke, vendar skuša čim bolj celovito zajeti kompleksnost terminoloških besednih zvez, za namene tega eksperimenta pa je bilo luščenje omejeno le na dva oblikoskladenjska vzorca, in sicer enega enobesednega (*samostalnik*) ter enega dvobesednega (*pridevnik* + *samostalnik*). Nanju smo se omejili zato, ker za enobesedne *samostalniške* termine ter termine v obliki zveze *pridevnik* + *samostalnik* velja, da so v slovenskem jeziku strokovnopoimenovalno najbolj produktivni (Logar Berginc et al. 2013). Oblikoskladenjske vzorce smo poleg tega omejili zgolj na občne samostalniške zveze brez lastnoimenskih sestavin, iz nadaljnje obravnave pa smo izločili tudi vse izraze, ki so se v gradivnih korpusih pojavili manj kot petkrat.

2.2 Korpus

Terminološke kandidate smo luščili iz dveh korpusov, in sicer iz:

a) specializiranega korpusa glasbenih besedil (dalje glasbeni korpus) in

b) podkorpusa ccGigafide, v katerega smo združili besedila s temami iz biologije, kemije in fizike (dalje naravoslovni podkorpus).

Glasbeni korpus obsega deset učbenikov, ki so nastali v obdobju desetih let (2004–2014) ter se uporabljajo pri pouku glasbe na osnovni in srednji stopnji glasbenih šol. Korpus je nastal v okviru doktorske raziskave Jelene Grazio na Oddelku za

muzikologijo Filozofske fakultete Univerze v Ljubljani, učbeniki pa obravnavajo različne glasbene prvine: harmonijo (učbenika Harmonija I in Harmonija II Janeza Osredkarja), kontrapunkt (Kontrapunkt Janeza Osredkarja), oblikoslovje (Oblikoslovje Larise Vrhunc), solfeggio (Solfeggio I, II, III in IV Tomaža Habeta) ter teorijo glasbe v splošnih potezah (Sodobna teorija glasbe P. Amaliettija in Osnove glasbene teorije Pavla Mihelčiča).

Naravoslovni podkorpus je sestavljen izključno iz besedil, ki so zajeta v ccGigafidi (Logar Berginc et al. 2012: 77–97; Erjavec in Logar Berginc 2012). V njem je 13 osnovno- in srednješolskih učbenikov za naravoslovje, biologijo, fiziko ali kemijo ter še 16 drugih strokovnih in poljudnostrokovnih knjig različnih založb na temo astronomije, botanike in vrtnarjenja. Ostala besedila so iz naslednjih revij: Gaia, Gea, Kmetovalec, Moj lepi vrt, Moj mali svet, Mrgolazen, National Geographic, Revija o konjih, Ribič ter Rože in vrt.[3]

Osnovne podatke o obeh podkorpusih povzema Tabela 1.

Tabela 1: Osnovni podatki o glasbenem korpusu in naravoslovnem podkorpusu.

	Glasbeni korpus	Naravoslovni podkorpus
število pojavnic	280.060	1.053.897
število različnic	12.121	59.788
število dokumentov	10	388
vrsta besedil	učbeniki	učbeniki, poljudnostrokovne knjige in revije

3 ANALIZA REZULTATOV

Število izluščenih terminoloških kandidatov za posamezni korpus in oblikoskladenjski vzorec kaže Tabela 2, v Tabelah 3 in 4 pa so podani vrhnji deli vseh štirih seznamov.

Tabela 2: Število izluščenih kandidatov za posamezni korpus in oblikoskladenjski vzorec.

	Glasbeni korpus	Naravoslovni podkorpus
samostalnik	1.137	7.853
pridevnik + samostalnik	828	1.309

3 Področja biologija, kemija in fizika v Gigafidi z besedili niso enakovredno zastopana, kar je razvidno tudi iz izluščenih terminoloških kandidatov.

Iz Tabele 2 je razvidna razlika med obema podkorpusoma v velikosti, ki pri samostalniških kandidatih izkazuje podobno razmerje med potencialno terminološkimi samostalniki in vsemi različnicami (9,3 % pri glasbi in 13 % pri naravoslovju), pri dvobesednih izrazih pa prednjači glasba s 6,8 % proti naravoslovju z 2,2 %.

Tabela 3: Vrhnji del obeh seznamov izluščenih terminoloških kandidatov iz glasbenega korpusa.

Samostalnik	Pogostost v Gigafidi	Pridevnik + samostalnik	Pogostost v Gigafidi
1. ton	32.538	osnovni ton	201
2. akord	3.049	pesemska oblika	17
3. oblika	357.909	vodilni ton	13
4. glasba	290.529	sonatna oblika	14
5. interval	6.620	tonovski način	24
6. stavek	43.996	taktovski način	10
7. tema	231.129	alterirani akord	0
8. tonaliteta	305	akordično območje	0
9. takt	6.515	osnovna tonaliteta	2
10. kvinta	189	notna vrednost	15
11. glas	272.023	durova lestvica	13
12. nota	29.362	alikvotni ton	5
13. terca	439	zgornji glas	9
14. dur	4.037	cel ton	14
15. oktava	603	stranska stopnja	0
16. melodija	27.394	tripolovinski takt	0
17. stopnja	199.417	glasbena teorija	269
18. lestvica	138.848	alterirani ton	2
19. septima	35	menjalni ton	1
20. način	551.177	tonski način	56
21. trozvok	72	osnovna oblika	1.036
22. skladba	81.732	notno črtovje	238
23. primer	906.970	molova lestvica	17
24. gibanje	136531	dominantni septakord	12
25. d	61.852	uvajalna vaja	7
26. četverozvok	10	dominantni četverozvok	0
27. c	42.002	velika terca	16
28. vaja	92.132	lestvična stopnja	0
29. polovinka	101	akordični ton	1
30. kadenca	356	tonalni plan	1

Tabela 4: Vrhnji del obeh seznamov izluščenih terminoloških kandidatov iz naravoslovnega podkorpusa.

Samostalnik	Pogostost v Gigafidi	Pridevnik + samostalnik	Pogostost v Gigafidi
1. rastlina	132.625	ribiška družina	5.599
2. voda	516.116	ribolovna dovolilnica	962
3. vrt	144.814	sladkovodno ribištvo	1.134
4. RDA	11.876	živa meja	5.147
5. riba	101.596	ribiška zveza	1.185
6. vrsta	612.575	organsko gnojilo	1.913
7. konj	96.764	botanični vrt	3.886
8. ribič	34.728	velika količina	25.964
9. leto	4695.764	okrasna rastlina	2.899
10. tla	162.784	mlad ribič	1.257
11. cvet	47.693	ribiški čuvaj	802
12. list	197.567	veliko število	52.554
13. sorta	45.117	zelenjavni vrt	2.337
14. žival	202.023	različna vrsta	14.756
15. čas	1950.895	ribiški okoliš	993
16. cm	104.836	soška postrv	1.595
17. slika	570.703	sadno drevje	4.192
18. delo	1703.484	ribji živelj	713
19. ribolov	19.679	hlevski gnoj	4.530
20. zemlja	161.980	potočna postrv	1.212
21. kg	111.327	cerkniško jezero	3.431
22. barva	245.458	cvetlični lonček	1.412
23. drevo	84.742	športni ribolov	1.893
24. površina	152.514	organska snov	2.915
25. oblika	368.503	nizka temperatura	9.432
26. seme	42.020	nova sorta	1.615
27. jezero	88.267	sladkorna pesa	3.549
28. temperatura	110.203	ekološko kmetovanje	3.097
29. prostor	691.860	visoka temperatura	10.700
30. fotografija	237.996	članska izkaznica	1.337

Pogled na Tabeli 3 in 4 pokaže dve temeljni razliki med podkorpusoma, in sicer v ravni specializiranosti in homogenosti. Predvsem pri seznamu izluščenih izrazov iz naravoslovnega korpusa opazimo vpliv različnih (pre)močno zastopanih virov, denimo s področij ribištva in vrtnarstva. Poleg tega so očitne tudi velike razlike v

pogostosti v Gigafidi, še posebej pri dvobesednih izrazih; bolj specializirani glasbeni korpus vsebuje izraze, ki se v Gigafidi redkeje pojavljajo kot naravoslovni izrazi, ki so bili izluščeni iz podkorpusa Gigafide.

V postopku analize smo natančno pregledali le vrhnjih 150 enot na seznamih terminoloških kandidatov iz obeh korpusov. Analiza je potrdila pričakovano razliko v številu enot, ki bi jih iz prvega oz. drugega korpusa dali v splošno košarico. V obeh seznamih iz glasbenega korpusa je bilo primerov za splošno košarico veliko manj kot v seznamih iz naravoslovnega podkorpusa, natančneje: samostalniških kandidatov, ki bi jih zelo verjetno umestili v splošno košarico,[4] je v glasbenem korpusu v vrhnjem delu seznama le približno tretjina (*takt*, *glas*, *nota*, *melodija*, *skladba*, *harmonija* ipd.), pri zvezah pridevnika in samostalnika pa celo manj kot 15 % (npr. *notno črtovje*, *klasična glasba*, *klavirska spremljava*), medtem ko bi ostale izluščene enote iz glasbenega korpusa sodile v šolsko ali strokovno (npr. v strokovno: *modulacija*, *fuga*, *kvintakord*, *tritonusna kvinta*, *eolska septima*, *napolitanski sekstakord*).[5]

Na drugi strani je stanje pri seznamih iz naravoslovnega podkorpusa obratno: samostalnikov in zvez iz Gigafidinega podkorpusa, ki bi jih dali v splošno košarico, je velika večina, prim.: *rastlina*, *voda*, *list*, *seme*, *plod*, *poganjek*, *temperatura*, *svetloba*; *okrasna rastlina*, *soška postrv*, *organski odpadek*, *listna uš* itd. To pomeni, da v vrhnjem delu seznama izluščenih enot iz naravoslovnega podkorpusa skoraj ni leksike, ki bi jo bilo treba obravnavati ožje terminološko, sploh pa ne v smislu strokovne košarice. Manjši del enot iz Gigafide bi tako lahko šel le še v šolsko košarico – izmed prvih 300 enobesednih smo sem uvrstili besede: *celica*, *molekula*, *muha* ('ribištvo'), *beljakovina*, *dušik*, *masa*, *spojina*, *atom*, *kromosom*, *sila*, *populacija* in *pH*. Če s pregledom seznama nadaljujemo do 500. mesta, bi sem lahko prišteli še poimenovanja: *križanec*, *bala*, *bakterija*, *podtaknjenec*, *ozvezdje*, *aminokislina*, *membrana*, *humus*, *elektron*, *kasač*, *uplenitelj*, *herbicid*, *gen*, *insekticid* in *siliranje*. Groba ocena torej kaže, da je v seznamu izluščenih samostalniških terminoloških kandidatov okrog 5-odstotni delež poimenovanj, ki imajo potencial za obravnavo v šolski košarici. Med večbesednimi kandidati je takih več, tj. okrog 15 % med vrhnjimi 300 enotami, npr. *ogljikov hidrat*, *celična membrana*, *magnetno polje*, *maščobna kislina*, *potencialna energija*, *vrtilni moment*.

Analiza je torej dala naslednje odgovore na naši prvi dve raziskovalni vprašanji:

1. Metoda luščenja terminoloških kandidatov v ccGigafidi praktično ne izkaže gradiva, ki bi zahtevalo ozko terminološko obravnavo, izkaže pa tudi zelo malo gradiva, pri katerem bi morali leksikografi paziti na

4 Podajamo subjektivno oceno, ki zadošča za preliminarne ugotovitve, pred dokončnimi zaključki pa bi jo bilo treba potrditi še z večjim številom ocenjevalcev.

5 Ne bi jih seveda dejansko umeščali v slovar, v mislih imamo le primerjavo med seznami.

področno vezanost iztočnice (morda tudi oznako). Gigafida torej – vsaj pri naravoslovnih besedilih in v vrhnjem delu seznama – v veliki večini vsebuje le poimenovanja, pri katerih, kot smo navedli že zgoraj, sicer prepoznamo navezavo na določeno strokovno področje, vendar poznavanje področja ni pogoj za njeno razumevanje in pravilno rabo.

2. V nasprotju s preteklimi luščenji, ki smo jih izvedli na specializiranih korpusih strokovnih besedil, smo tokrat metodo preizkusili na področno heterogenem naboru poljudnostrokovnih besedil. Taka so namreč tipična besedila v splošnih korpusih: ubesedujejo teme, ki povezujejo različna področja, ter jih opisujejo in razlagajo na nestrokovnjakom prilagojen način. Eksperiment je pokazal, da je z vidika analize rezultatov prihodnja taka luščenja bolje načrtovati na tematsko kolikor se da enotnih besedilnih zbirkah, četudi bi se na ta način opredeljeni viri (npr. revija Gea) pojavljali v več izvedbah te metode. Za leksikografsko analizo je namreč manj moteče, če smo pri pregledu osredotočeni na poimenovanja (in njihovo ožjo terminološkost) zgolj enega področja. Obenem je treba poudariti, da bi za namene slovarja, ki naj bi zajemal tudi specializirano izrazje strok na ravni srednješolskih učbenikov, splošni korpus morali dopolniti z ustreznimi učbeniki, pri luščenju pa uporabiti primerjavo med homogenim specializiranim korpusom in splošnim korpusom brez stvarnih besedil.
3. Razvrščanje v košarice mora v končni implementaciji luščenja potekati samodejno. Čeprav se že v tu opisanem pilotnem eksperimentu zarisujejo frekvenčna območja, v katerih se gibljejo izrazi v splošni, šolski in strokovni košarici, je treba metodologijo razvrščanja šele razviti skupaj s spremenljivkami, ki bodo odvisne od posameznega strokovnega področja (npr. frekvenčni pragovi in oblikoskladenjski vzorci za luščenje). Namesto absolutne korpusne pogostosti bo metoda po vsej verjetnosti uporabljala pogostostno razmerje med strokovnim in splošnim delom korpusa oziroma navzkrižne primerjave med podkorpusi sorodnih področij. Temeljna ovira, da takšne metodologije v tem trenutku še ni mogoče predlagati, je neobstoj orodja za samodejno razdvoumljanje, ki bi omogočalo razdelitev korpusne pojavitve izrazov na posamezne pomene, šele s dostopnostjo podatka o pogostosti posameznega pomena pa lahko učinkovito (in tudi samodejno) presojamo o terminološkosti.

Ker je pri sedanji metodi luščenja za izračun terminološkosti uporabljena primerjava pogostosti med specializiranim in celotnim splošnim korpusom, se pri rezultatih jasno izkaže tudi nesorazmerna zastopanost področij v Gigafidi; tako se denimo pogostosti določenih leksikalnih enot lahko nesorazmerno povečajo zgolj zaradi določene revije, ki je v korpus vključena z več letniki. Čeprav se načrtuje

sistematično dopolnjevanje Gigafide z besedili tistih področij, ki so zdaj slabše zastopana, bi bilo iluzorno pričakovati, da bo uravnoteženost korpusa kdajkoli popolna; pravzaprav področna uravnoteženost niti ni cilj referenčnega korpusa.

4 SKLEP

Ob koncu velja še enkrat poudariti, da nas v eksperimentu ni zanimala ocena terminološke uspešnosti luščenj iz dveh korpusov – enega področno specializiranega in homogenega, drugega splošnejšega in heterogenega – v smislu terminografije, temveč smo si zadali nalogo s to metodo poiskati terminološkost v splošnem jeziku. Predvidevali smo, da bi preliminarni rezultati lahko izpostavili nekatere prednosti in slabosti ter nakazali nekatere smernice za prilagoditev pristopa, ki ima prvotno drugačen namen. Čeprav smo iz luščenja izpustili časopisje, internetna besedila in kategorijo »drugo«,[6] so seznami, pridobljeni po trenutni metodi luščenja iz heterogenega korpusa naravoslovja, izkazali majhen obseg enot, ki bi v prihodnjem slovarju slovenščine, nastalem na podlagi Gigafide, potrebovale terminološko pozornost, nasprotno sliko pa kaže glasbeni korpus. Analiza seznamov je pokazala tudi to, da bo za leksikografski postopek potrebna še nadaljnja natančnejša opredelitev košaric, pri nadgradnji korpusa pa temeljitejša dopolnitev z besedili, ki vsebujejo terminologijo, s katero se v času izobraževanja sreča velika večina govorcev slovenščine, tj. z učbeniki in sorodnim gradivom (Logar 2015).

6 Ta del Gigafide prav tako vsebuje determinologizirano leksiko, prim. npr. pravna poimenovanja *pogodba*, *odločba*, *sklep*, *pritožba*, *zahtevek*, značilna za spletni del korpusa, v Logar Berginc in Ljubešić (2013: 102).

Analiza iskalnih poizvedb na portalu Termania

Špela Vintar

Abstract
This paper describes an analysis of search query logs for the Termania.net dictionary portal. The aim of the analysis is to shed light on user information needs regarding specialised vocabulary. A brief overview of related approaches is given, although none focus specifically on specialised terms. The list of queries to Termania.net is compared to the frequency lists of the Gigafida and the EnTenTen corpora, and the list of entries in two editions of the current dictionary of Slovenian: the SSKJ and SSKJ2. In accordance with related recent findings, users of Termania often search for frequent Slovenian words, but a large portion of the queries is also aimed at specialised vocabulary. The analysis leads to tentative conclusions about the future role of a dictionary of contemporary Slovenian regarding specialised vocabulary.

Keywords: Termania, search queries, dictionary users, specialised vocabulary

Ključne besede: Termania, iskalne poizvedbe, slovarski uporabniki, specializirana leksika

1 UVOD

V uvodnem prispevku sklopa o specializirani leksiki smo že namenili nekaj pozornosti uporabniškim potrebam, med drugim tudi tujim študijam, ki skušajo z uporabniškega vidika opredeliti vlogo specializirane leksike v splošnem slovarju (Josseline-Leray in Roberts 2005; Müller-Spitzer 2014). V tem prispevku opisujemo analizo dnevnika iskalnih poizvedb na portalu Termania,[1] s katero želimo odstreti še en delček v kompleksnem mozaiku uporabniških profilov, njihovih informacijskih potreb in spletnih navad.

Analize dnevnikov iskanj so za raziskovanje potreb in navad uporabnikov slovarskih spletišč uporabili že številni avtorji. Kratek pregled tovrstnih raziskav podaja Lew (2015), ki med uspešnejšimi zgodnjimi študijami izpostavlja Lemnitzerja (2001, citirano po Lew 2015), saj je bil neposredni rezultat analiz uporabniških poizvedb izboljšan spletni vmesnik in vgradnja novih iskalnih funkcij. Bergenholtz in Johnsen (2005) opisujeta analizo dnevnikov iskanj na danskem slovarskem portalu in navajata več koristnih opažanj, denimo da so med nenajdenimi izrazi v veliki večini zatipkane besede in da bi si uporabniki danskega slovarja močno želeli iskati glagolske oblike v trpniku, a iskalnik tega ne omogoča. Lorentzen in Thailgaard (2012) primerjata iskalne poizvedbe, ki uporabnike pripeljejo do slovarskega portala, s poizvedbami, ki jih uporabniki vpišejo v iskalnik na portalu. Zanimiva ugotovitev je predvsem razlika v pojavnosti večbesednih poizvedb: medtem ko je pri spletnih iskalnikih takšnih poizvedb 40 odstotkov, je bilo na slovarskem portalu kar 96 odstotkov poizvedb enobesednih.

Lew (2015) priznava, da so dnevniki iskanj sicer praktičen vir podatkov o slovarskih uporabnikih, saj jih običajno beležijo že strežniki in se raziskovalcem - za razliko od ostalih metod uporabniških študij - zanje ni treba posebej truditi, vendar po drugi strani dajejo precej okrnjeno podobo uporabniške situacije, kajti iz seznama poizvedb ne izvemo nič o kontekstu iskanja ter o vzrokih in posledicah posameznega iskalnega dejanja. Prav tako dnevniki iskanj ne vsebujejo nikakršnih podatkov o uporabniku, kot so npr. starost, spol, poklic ali jezikovna kompetenca. Ena sodobnejših raziskav (Hult 2012) uporablja bolj kompleksno metodo, ki dnevnike iskanj povezuje z vprašalniki, oboje skupaj pa nam prek številke IP omogoča identifikacijo določenega uporabnika in natančno spremljanje celotne seje na slovarskem portalu, pa tudi analizo ponovnih obiskov iste osebe in podrobno profiliranje njenega iskalnega vedenja.

Dnevnike iskanj lahko uporabljamo tudi za vpogled v pogosto, redko ali nenajdeno iskano besedišče. Tako si de Schryver et al. (2006) postavljajo vprašanje, ali je korpusna pogostost zares dober kriterij za vključevanje leksemov v slovar, in s

1 http://www.termania.net (dostop 8. 8. 2015).

pomočjo analize dnevnikov iskanj ugotovijo, da je pomen korpusne pogostosti precenjen. Koplenig et al. (2014) so uporabili drugačno metodologijo primerjave seznama poizvedb s pogostostnim seznamom korpusa in prišli do nasprotnih ugotovitev, in sicer da uporabniki iščejo tudi pogoste besede oziroma da korelacija med korpusno pogostostjo in pogostostjo iskanj obstaja. Nobena od tu omenjenih študij in nam znanih raziskav se ni ukvarjala z analizo dnevnikov iskanj z vidika specializiranega izrazja.

2 UPORABLJENA METODOLOGIJA

Če bi prosili prevajalca, naj našteje najpomembnejše in najobsežnejše spletne vire terminologije v Sloveniji, bi se na seznamu verjetno na prvem mestu znašla Termania, sledili pa bi ji še Evroterm,[2] Islovar[3] in Terminologišče.[4] To nikakor ni izčrpen seznam, saj denimo na skrbno vzdrževanem spletnem imeniku Prosto dostopni slovarji,[5] za katerega skrbi Miran Željko, najdemo povezave do prek 1.100 spletnih slovarjev, od tega prek 200 specializiranih eno-ali večjezičnih slovarjev slovenskega izrazja najrazličnejših področij.

Kosem (2014) večslovarska spletišča deli na portale, agregatorje in zbirke slovarskih povezav. V skladu s to tipologijo bi med portale lahko uvrstili Terminologišče[6] in FRAN,[7] kjer so na voljo slovarji enega založnika, tj. Založbe ZRC SAZU, Termanio pa med slovarske agregatorje, saj ponuja iskanje po številnih slovarjih iz različnih virov prek enotnega iskalnega vmesnika. Evroterm in Islovar sta obsežni spletno dostopni terminološki bazi, ki sta (bili vsaj izvorno) specializirani za izrazje v zvezi z EU oziroma računalniško in informatično izrazje.

Za ugotavljanje terminoloških potreb slovenskih uporabnikov bi bilo zanimivo analizirati strežniške dnevnike vseh pomembnejših slovarskih spletnih mest, a to iz različnih razlogov ni bilo mogoče. Od podjetja Amebis smo pridobili seznam poizvedb za Termanio za pretekli dve leti in pol, prav tako nam je Slovensko društvo Informatika brez zadržkov posredovalo seznam najpogostejših poizvedb in nenajdenih izrazov v Islovarju, a slednjega žal v formatu, ki ni primeren za količinsko obdelavo. Baza Evroterm, ki vsebuje bogato zbirko pravnega izrazja ter izrazja vseh področij, ki se navezujejo na EU, presenetljivo nima urejenega arhiviranja strežniških poizvedb. Pri pričujoči raziskavi smo se tako omejili na

2 http://www.evroterm.gov.si/ (dostop 8. 8. 2015).
3 http://www.islovar.org (dostop 8. 8. 2015).
4 http://isjfr.zrc-sazu.si/en/terminologisce#v (dostop 8. 8. 2015).
5 http://www.evroterm.gov.si/slovar/index.html (dostop 8. 8. 2015).
6 Terminologišče pravzaprav ni pravi portal, kajti iskanje je mogoče le po vsakem slovarju posebej; to težavo delno rešuje portal Fran, ki prav tako vključuje terminološke slovarje, vendar v primeru zadetka ponudi le povezavo na ustrezni slovar v Terminologišču.
7 http://www.fran.si/ (dostop 8. 8. 2015).

Termanio, ki sodeč po številu poizvedb sodi med najbolj obiskana slovarska spletišča pri nas, hkrati pa je zaradi vsebovanih terminoloških virov zanimiva tudi z vidika specializiranega izrazja.

Podjetje Amebis je za potrebe pričujoče analize prispevalo seznam iskalnih poizvedb, opremljenih s pogostostmi. Seznam je izdelan na podlagi strežniških dnevnikov za zadnji dve leti in pol in skupno obsega 433.692 različnih iskalnih poizvedb, v skupnem seštevku pa skoraj 6 milijonov in pol poizvedb. 287.283 poizvedb je bilo vpisanih le enkrat. Žal iz seznama ni razvidno, katere poizvedbe so bile uspešne in katere ne. Prav tako seznam ne vsebuje drugih podatkov, kot so datumi in ure poizvedb ali številke IP, s katerih so bile poslane poizvedbe.

Pri analizi nas je zanimalo predvsem, ali uporabniki na Termanii iščejo tudi specializirane izraze, ob tem pa smo želeli pridobiti tudi splošnejši vtis o tipih iskanj, iz katerih bi morda lahko sklepali o informacijskih potrebah uporabnikov. Poleg ročnega pregleda najpogostejših iskanj smo celotnemu seznamu avtomatsko pripisali še naslednje podatke:

- pogostost v korpusu Gigafida (GF),
- prisotnost v geslovniku Slovarja slovenskega knjižnega jezika prve izdaje (SSKJ1),
- prisotnost v geslovniku Slovarja slovenskega knjižnega jezika druge izdaje (SSKJ2) in
- pogostost v korpusu angleškega jezika EnTenTen (EN1010) (Jakubiček in dr. 2013).

S slednjim podatkom smo želeli predvsem identificirati in izločiti angleške izraze, ki so se med poizvedbami pogosto pojavljali, ob tem pa velja pripomniti, da je veliko poizvedb tudi po nemških, italijanskih, francoskih in drugih iztočnicah, ki jih nismo samodejno izločali.

3 TERMANIA

Termania.net je slovarski agregator, ki ga je leta 2010 vzpostavilo podjetje Amebis d. o. o. kot enotno in brezplačno vstopno točko do številnih eno- in večjezičnih, splošnih in specializiranih, slovenskih in tujih slovarskih zbirk. Za slovenščino je prek portala mogoče iskati po 44 slovarjih, od katerih jih je polovica uvrščenih v kategorijo splošnih virov. Poleg SSKJ in Slovenskega pravopisa tako med splošnimi najdemo še Leksikalno bazo slovenščine, Sloleks, slovarje rim, okrajšav, slengov in dialektov, vezljivostni slovar ter številne splošne dvojezične slovarje v kombinaciji s slovenščino. Od terminoloških virov so vključeni manjši in večji slovarji različnih

strok, od izobraževanja, informacijske znanosti, računalništva, elektrotehnike in medicine do fotografije, rokodelstva in klekljarstva (glej Sliko 1).

Iskanje je na voljo v enostavni in napredni različici. Enostavno iskanje sproži poizvedbo po vseh slovarjih, zadetki pa se prikažejo v obliki skrajšanih gesel. Ker pri enostavnem iskanju poizvedba lahko prikliče veliko število zadetkov iz različnih slovarjev, se za razvrščanje rezultatov uporabijo različni kriteriji, kot so izbrani jezik vmesnika, mesto zadetka v geselskem članku, pomembnost slovarja, abecedni vrstni red (Romih in Krek 2012). S pomočjo naprednega iskanja ima uporabnik možnost omejiti iskanje na določeno lastnost, kot je element slovarja (iztočnica, prevod, drugo), jezik, področje ali slovar. Tako pri osnovnem kot pri naprednem iskanju lahko poleg iskanja ene besede iščemo tudi po več besedah (besednih zvezah), uporabljamo pa lahko tudi posebne znake (* in ?), s pomočjo katerih lahko še dodatno razširimo ali omejimo iskanje.

Čeprav bi morda iz imena portala sklepali drugače, je Termania kljub širokemu naboru terminoloških slovarjev pretežno splošni slovarski portal, ki po vsej verjetnosti mnogim uporabnikom ustreza prav zato, ker na enem mestu ponuja SSKJ in dvojezične slovarje. Za izpolnjevanje na uvodni strani zastavljenega cilja, da bi portal združeval vse pomembnejše vire terminologije za slovenščino, bi bilo vanj treba integrirati vsaj še terminološke slovarje ZRC SAZU in Evroterm.

Slika 1: Seznam področij, ki so vključena v Termanio.

4 ANALIZA SEZNAMA POIZVEDB NA TERMANII

V nadaljevanju predstavljamo rezultate analize seznama poizvedb. Ko smo seznam avtomatsko opremili s pogostostmi v korpusih Gigafida in EN1010 ter s podatkom o vključenosti v SSKJ oziroma SSKJ2, smo lahko z enostavnim filtriranjem seznama v Excelu opazovali različne presečne množice, pa tudi sezname poizvedb, ki jih ni v nobenem od primerjanih virov.

Med najpogostejšimi poizvedbami (glej Tabelo 1) je večina splošnih slovenskih besed, ki se pojavljajo tudi v GF, SSKJ in SSKJ2, npr. *hiša, ki, vrsta, del, oblika, zveza, lastnost, poročilo*. Sledijo angleške besede, ki so z veliko pogostostjo izpričane v EN1010, z majhno tudi v GF, v slovenskih slovarjih pa jih ni: *control, system, check, note, call, business*. Prav na vrhu seznama je tudi nekaj poizvedb, ki jih ne moremo z gotovostjo umestiti v določen jezik, npr. *A, Ga, n., SEM* in *GL*. Nekatere bi se utegnile nanašati na kratice in okrajšave. Prav tako pri nekaterih poizvedbah ne moremo vedeti, ali so ciljale na slovensko, angleško ali kako drugo iztočnico; tak primer je izraz *rod* na 11. mestu. Vsaj kar se tiče slovenskih besed, se delno potrjujejo ugotovitve Koplenig et al. (2014), ki so na primeru poizvedovanj na nemškem slovarskem portalu DWDS in nemškem Wictionaryju ugotovili, da uporabniki pogosto iščejo besede, ki so pogoste tudi v referenčnem korpusu.

Tabela 1: Najpogostejša iskanja na Termanii in primerjava z Gigafido, SSKJ, SSKJ2 in EN1010.

	Termania	GF	SSKJ	SSKJ2	EN1010
hiša	239.206	484.014	da	da	0
A	234.390	6.677	ne	ne	389.126
ki	185.690	11.669.026	da	da	237
Ga	126.139	3	ne	ne	16.217
vrsta	85.713	577.698	da	da	0
n.	84.090	26.754	ne	ne	5.272
del	72.118	1.477.271	da	da	21.573
oblika	67.153	357.909	da	da	0
SEM	51.716	12	ne	ne	2.630
GL	40.158	2.327	ne	ne	3.115
rod	34.806	58.862	da	da	31.441
control	24.951	1.971	ne	ne	927.064
zveza	24.505	494.360	da	da	0
lastnost	23.955	94.964	da	da	0
poročilo	22.863	328.600	da	da	0
system	20.903	1.612	ne	ne	1.882.460
check	20.046	1.841	ne	ne	505.989
ugotoviti	19.896	310.831	da	da	0
značilnost	18.899	57.441	da	da	0

Nadaljnja primerjava seznama poizvedb s preostalimi štirimi viri prikaže pestro podobo iskanj (Tabela 2). Najprej ugotovimo, da je prek četrtino poizvedb moč najti v korpusu angleškega jezika, iz česar lahko sklepamo, da se Termania pogosto

uporablja kot dvojezični portal. Dobrih štirideset odstotkov iskanj se vsaj enkrat pojavi tudi v korpusu Gigafida, a ker smo primerjavo opravili s seznamom lem iz Gigafide, je med poizvedbami na Termanii v resnici še precej več slovenskih besed, ki so jih uporabniki zapisali v sklanjani ali spregani obliki. Zgolj 14 odstotkov različnih poizvedb je obrodilo zadetke iz SSKJ, vendar je skupno število teh poizvedb približno 2,6 milijona oziroma 40,6 odstotkov vseh poizvedb.

Tabela 2: Številčna primerjava seznama poizvedb s korpusoma GF in EN1010 ter slovarjema SSKJ in SSKJ2.

	Število poizvedb	Odstotek poizvedb
Termania	**433.692**	**100 %**
Termania - 1x	287.283	66,3 %
je v GF	177.687	40,1 %
je v SSKJ	61.393	14 %
je v SSKJ2	64.348	15 %
je v EN1010	114.044	26 %

Zgolj 91 poizvedb najdemo v SSKJ2, ne pa v SSKJ; ni presenetljivo, da je med njimi večina novejših izrazov, tudi terminoloških: *spazmolitik, antiemetik, telemedicina, bioptičen, terminologizacija, všečkati, e-pošta, prostovoljiti, e-naslov, maskarpone, razpoznavalnik, tviteraš.*

Poizvedbe, ki jih ni v nobenem od primerjanih virov, bi lahko glede na pogostost razdelili v naslednje kategorije:

- slovenske besede v neosnovni obliki, pretežno neterminološke: *podatki, iščem, priljubljena, spremljaj, zadržano;*
- angleške besede v neosnovni obliki: *prospecting, sieving, levying, inventorying, garnishing;*
- slovenski terminološki izrazi, pretežno tujke: *hipersoničen, ekhimoza, acianotičen, transudat, distenzija, mezotelij, hipersplenizem;*
- iskanja z zvezdico: *an*, k*, turist*, pos*, spoln*, hidroksi*;*
- iskanja končnic z vezajem (takšnih iskanj iskalnik Termanie ne podpira in so zato vedno neuspešna): *-olg, -okate, -njiva, -njice;*
- kratice, krajšave, akronimi: *crh, ToR, ZAZV, Tfc, accn;*
- tujejezične besede razen angleških: *einrichtung, belegen, verteilen, stellung, ausgleich, Spannungsversorgung, knikken, gebruiken, plutajuči, το θηρζον;*
- ostalo: lastna imena, besede v oklepajih ali navednicah, zatipkanke, neprepoznavni nesmisli, števila in simboli.

Kar 94.891 poizvedb je večbesednih, pri čemer so najpogostejše dvobesedne, kar nekaj pa je tudi primerov, ko so uporabniki v iskalno polje prilepili kar cel odstavek besedila. Najpogostejši večbesedni iskalni niz je *severna amerika* (2833), sledijo *jugovzhodna azija* (481), *vzporedna izdaja* (469), *vrstniško nasilje* (363) in *mazivno olje* (341). V nadaljevanju med večbesednimi poizvedbami najdemo ogromno terminoloških: *in vitro* (76), *vena cava* (45), *retrogradna pielografija* (42), *violinski ključ* (35), *diabetes mellitus* (34), *proceduralno znanje* (26), *multipla skleroza* (24); pa tudi večbesednih angleških izrazov: *denim trousers* (46), *fava bean* (44), *time lapse* (33), *domain entity* (33), *stem cell* (27), *third party* (26), *according to* (26).

5 RAZPRAVA IN ZAKLJUČKI

Poizvedbe so sledi uporabnikov, iz katerih lahko sklepamo o njihovih informacijskih potrebah, namerah in navadah, vendar se moramo ob tem zavedati omejitev tovrstnih analiz. Tako iz dnevnika iskanj ni razvidno, ali so uporabniki iskali prek osnovnega ali naprednega iskanja, ali so iskanje omejili na določeno skupino slovarjev, ali je iskanje vrnilo zadetke in ali so bile pridobljene informacije zanje uporabne. Prav tako se ne moremo povsem zanašati na pogostosti poizvedb, kajti nekateri brskalniki isto poizvedbo pošljejo večkrat, različne aplikacije - denimo prevajalski programi - pa uporabljajo določene nize za testiranje dostopa ali poizvedujejo samodejno. Kljub tem omejitvam pa menimo, da je iz Termaniinih dnevnikov mogoče razbrati, kaj si uporabniki želijo od slovarskega portala, še posebej, če ta obljublja »vsemogočnost« v smislu združevanja najrazličnejših slovarskih zbirk v skupno iskalno okolje.

Tako naša analiza kaže, da ima približno 40 odstotkov vseh poizvedb (kumulativno) zadetek iz SSKJ, torej gre pri teh poizvedbah za slovenske izraze splošnega ali nizko- do srednjespecializiranega besedišča.[8] Verjetno je dobršen delež teh poizvedb v resnici usmerjen v večjezične slovarje, v katerih so želeli uporabniki, domnevno učenci, študenti, pedagogi in drugi, poiskati tujejezični ustreznik iskanega izraza.

Naslednja velika skupina poizvedb je v angleščini, kar je v kontekstu tukajšnje raziskave manj relevantno, a kaže na akutno potrebo slovenskih uporabnikov po velikem prosto dostopnem angleško-slovenskem slovarju. Iz pogovorov s skrbniki drugih dveh velikih terminoloških portalov, Evroterma in Islovarja,[9] je razvidno,

8 Pri tej tezi se naslanjamo na metodološka izhodišča SSKJ glede terminologije, ki so podrobneje predstavljena v Vintar (2015).

9 Osebni pogovor z Adriano Krstič-Sedej, terminologinjo Evroterma (januar 2015), in Jurijem Jakličem ter Tomažem Turkom, urednikoma Islovarja (december 2014).

da uporabniki tudi na teh dveh spletiščih pogosto iščejo zelo splošne besede in želijo terminološko bazo uporabljati kot splošni dvojezični slovar.

Sledijo poizvedbe, ki so usmerjene v slovenske eno- in večbesedne srednje- in visokospecializirane izraze. Največ jih je vezano na področje medicine, najbrž tudi kot odraz dejstva, da je Slovenski medicinski slovar s svojimi 43.792 gesli najobsežnejši in najdragocenejši terminološki vir, ki ga Termania ponuja. Za razliko od obeh prej omenjenih skupin poizvedb tu domnevamo, da uporabnike zanima enojezična slovarska informacija, saj poleg medicinskih izrazov izstopajo še tujke, kratice in okrajšave, latinski izrazi ter tehnični in pravni izrazi. Med njimi so tudi poizvedbe, ki v SSKJ nimajo zadetka, a se v GF pojavljajo več kot tisočkrat, npr. *toplice, atopijski, IT, štartnina, šprinter, visokotehnološki*.

V tej heterogeni skupini delno specializiranih poizvedb vidimo prostor za nastajajoči slovar sodobne slovenščine, pri čemer naj bi ta izpolnjeval informacijske potrebe po nizko- in srednjespecializiranih izrazih, tudi tujkah, medtem ko bi za visokospecializirane izraze ter terminološke definicije potrebovali pravi terminološki portal. Ta bi moral resnično ponujati dostop do vseh obstoječih terminoloških zbirk, predvsem pa bi moral naslavljati ožjo in zahtevnejšo skupino uporabnikov.

Če sklenemo, se kljub vsem omejitvam predstavljene analize iz seznamov poizvedb vendarle izriše nekaj ugotovitev, ki se zdijo koristne pri snovanju novega slovarja sodobne slovenščine, skupaj z drugimi uporabniškimi raziskavami pa dajejo popolnejšo sliko o informacijskih potrebah in željah (bodočih) uporabnikov slovarja. Pri snovanju spletnih slovarskih vmesnikov bi zato veljalo posebno pozornost nameniti načinu beleženja poizvedb, kajti iz podrobnejših strežniških dnevnikov lahko z današnjo tehnologijo že precej natančno spremljamo tipe uporabnikov in njihove načine uporabe slovarskih spletišč.

VIII
Stilistika in jezikovna raba v slovarskem opisu

Stilistika in enojezični slovar: označevanje jezikovne variantnosti

Monika Kalin Golob in Polona Gantar

Abstract
This paper places style and usage labels within the framework of stylistics for the purpose of constructing a new dictionary of contemporary Slovenian. Following a summary of the history of stylistics as a linguistic discipline and differing views on its subject matter, an overview is provided of stylistic qualifiers in Slovenian dictionaries and two lexicographic concepts: the Proposal for a Dictionary of Contemporary Slovenian (2013) and the Draft Proposal for a New Dictionary of Literary Slovenian by the Fran Ramovš Institute of the Slovenian Language (2015). We conclude with a proposal for marking variance within language use. This concept is largely based on the communicative-informative principle developed for the Slovenian Lexical Database. We enhance this description by displaying a number of different possibilities for integrating complex stylistic information into different segments within a single dictionary entry.

Keywords: stylistics, lexicography, stylistic marking, lexical database, dictionary

Ključne besede: stilistika, leksikografija, stilistično označevanje, slovarska baza, slovar

1 UVOD

Stilistika kot naslednica grške retorike, poetike in dialogike od svoje disciplinarne osamosvojitve v drugi polovici 19. stoletja ter razvoja v jezikoslovno stilistiko, ki jo začenja Ballyjeva Stilistika francoskega jezika (1909), v 20. stoletju kot relativno samostojna in v 50. letih od literarne teorije osamosvojena jezikoslovna disciplina proučuje izrazna sredstva (jezikovna in izvenjezikovna), ki sodelujejo pri nastanku besedila, vprašanja zgradbe in strukture besedila ter stilno diferenciacijo jezikovnih pojavov (prim. Mistrík 1988: 7–30; Čechová et al. 1997: 11).

Zaznamovana z različnimi filozofskimi smermi 19. stoletja in različnimi jezikoslovnimi nazori v prvi polovici 20. stoletja (od ruskega formalizma, Saussurjevega strukturalizma, praškega funkcionalizma v Evropi, prek cambriške šole v Angliji in novega kriticizma v ZDA) je v drugi polovici 20. stoletja razvila zelo različne metodološke pristope. V Sovjetski zvezi se je ob vključevanju statističnih metod in z analiziranjem t. i. funkcijskih stilov raziskovanje usmerjalo k stilistiki besedila, s katero se metodološko eksaktno v 60. in 70. letih ukvarjajo v anglo-ameriškem svetu, tudi v povezavi s statističnimi metodami in razvojem pragmatične stilistike. Pomemben preokret od ruskega formalizma, ki so mu očitali omejenost zgolj na literarna besedila ter njihovo formo brez upoštevanja ostalih dejavnikov in ki je hkrati onemogočal analizo daljših besedil, v zahodnem svetu prinaša Hallidayev funkcijski model konec 70. let (1976). Z usmeritvijo v družbene okoliščine in kontekstualne dejavnike, kot so register, spol, ideologija, je Halliday vpival na tiste smeri stilistike, ki se usmerjajo na jezikovno manifestacijo ideologije, npr. feministično in kritično stilistiko (Norgaard et al. 2010: 3). Besediloslovna stilistika je v drugi polovici 20. stoletja razvita tudi v Nemčiji in Avstriji, v romansko govorečih državah pa so bile raziskave usmerjene k ekspresivni in individualistični stilistiki (Mistrík 1988: 23). Na slovensko jezikoslovje je imela od 60. let naprej velik vpliv češka stilistika, ki se je po utemeljitvi v 30. letih s Praškim lingvističnim krožkom (PLK) v 60. in 70. letih ukvarjala predvsem z raziskovanjem posameznih funkcijskih jezikovnih stilov (pri nas poimenovanih zvrsti) ter kasneje z vprašanji besedila, tudi v povezavi s sociolingvistiko (prim. Mistrík 1988: 25, 26; Čechová et al. 1997: 245 in nasl.). Čeprav je strukturalizem v stilistiko prinesel objektivizacijo raziskovalnih metod (predvsem v raziskovanje literarnih besedil), je popolnoma prezrl zunajjezikovne razsežnosti, kar poskušajo preseči poststrukturalistična raziskovanja, predvsem s preseganjem binarnega raziskovanja in usmerjanjem k večpomenskosti leksike, poudarjanjem vloge kognitivnih procesov in jezikovnih razmerij (Crystal 1998: 78, 79).

V 21. stoletju je stilistika še vedno izrazito interdisciplinarna (Hoffmanová 1997) oz. kar transdisciplinarna veda (Katnić-Bakaršić 2007: 46) s stičišči različnih teorij in pogledov na jezik v rabi. Zaradi raziskovanja jezika kot kontekstualiziranega

ostaja meddisciplinarno področje različnih disciplin jezikoslovja (Hoffmanová1997; Sorlin 2014: 6). V zahodnem jezikoslovju je sicer še vedno razumljena kot raziskovanje predvsem leposlovnih besedil (Crystal 1998: 70–79; Norgaard et al. 2010: 2[1]), kar pa se je v sodobnosti spremenilo z vključevanjem stilističnega raziskovanja oglaševalskih, strokovno-znanstvenih, novinarskih in drugih neumetnostnih besedil, čemur se je del vzhodnoevropske in z raziskovanjem posameznih »funkcijskih zvrsti« tudi slovenska stilistika posvečal že od 60. let 20. stoletja pod vplivom češke funkcijske stilistike.

Izhodišče strukturalističnega stilističnega raziskovanja je bila dihotomija med stilno zaznamovanostjo in nezaznamovanostjo jezikovnega pojava: kar je zaznamovano, ni nevtralno. A vendarle bipolarnost ni preprosta, ni enoumna, saj ob inherentni zaznamovanosti obstaja tudi adherentna, kontekstualno ustvarjena; ob sistemski variantnosti pa široka paleta različnih besedilnih variant, ki znotraj besedilne vrste ali določenih okoliščin rabe delujejo pričakovano (lahko bi rekli opazno nezaznamovano) in zato tvorijo tipično »stilno plast«, ki zunaj teh okoliščin postane opazna. Ničta točka, tj. točka, ki je izhodišče presojanja za razlikovanje med zaznamovanim in nezaznamovanim, namreč ni stalna, ampak spremenljiva in odvisna od pogleda na določen pojav, ki nikoli ni izoliran v praznem prostoru, ampak rezultat družbenih, kulturnih, jezikovnih okoliščin in pričakovanj. Prav zato je bil v poststrukturalistični stilistiki tak bipolarni pogled presežen in pod vplivom novih metod v jezikoslovje vključeni novi modeli raziskovanja, ki jih danes poznamo kot pragmatična (razvita sicer v 60. letih 20. stoletja, a pomembnejša od 90. let), korpusna in večznakovna (multimodalna) stilistika (prim. Norgaard et al. 2010).

Poststrukturalistični jezikoslovni pristopi, ki so lastni tudi kritični stilistiki in v marsičem skladni s postulati korpusnega jezikoslovja, so zato uporabni tudi za razmislek o načinu stilističnega označevanja v slovarjih, ki temeljijo na modelu jezikovnega opisa na podlagi korpusne analize. V okviru poststrukturalnega pristopa tako ne iščemo novih sistemizacij in tipologizacij, ampak težimo k opisu diskurzivnih praks, kar lahko tudi pri slovarskem opisu pomeni, da ne želimo vnaprej sistemizirati in tipologizirati, ampak opisati ter iz opisa zgraditi tudi nov pristop k leksikalni stilistiki.

Od grške retorike, ki je razlikovala med tremi stili (visoki, srednji in nizki) ter jim pripisovala tipične jezikovne uresničitve, do funkcijskih stilov PLK, ki je v vzhodnoevropskem in slovenskem jezikoslovju in stilistiki 20. stoletja postavil temelje razumevanja jezikovne razslojenosti, je osnovno stilistično raziskovanje povezano

1 »Stilistika je študij načinov, po katerih je ustvarjen pomen v leposlovnih, pa tudi drugih vrstah besedil. /.../ Stilistika je pogosto razumljena kot jezikoslovni pristop k leposlovju – kar je razumljivo, saj je bila večina stilistične pozornosti do sedaj posvečena leposlovnim besedilom.«

z možnostjo izbire jezikovnih sredstev, da bi dosegli sporočanjski namen, bodisi pričakovano/neopazno/nezaznamovano bodisi nepričakovano/opazno/zaznamovano (kar PLK definira s pojmoma jezikovne rabe kot avtomatizirane oziroma aktualizirane). Po Akhmanovi (1976) jezikoslovno stilistiko zanima tista vrednost, ki je jezikovnemu sredstvu dodana ob denotativnem pomenu, torej konotativni pomen, ki ga vsebuje inherentno zaznamovani leksem ob denotativnem pomenu ali adherentno s priklicem in v kontekstu ustvarjene dodane vrednosti jezikovnih sredstev.[2] Jezikovne enote so deli sporočanjskega sistema in njihova sporočanjski dodana vloga je po Akhmanovi predmet lingvostilistike. Zanimanje jezikoslovne stilistike sega danes dlje in zajema vsa besedila in diskurze, medtem ko Akhmanova pozornost posveča leposlovju in poetski funkciji ter sinonimiji kot pomembni kategoriji stilističnega raziskovanja.

Slovenska slovnica (Toporišič 2000: 13 in nasl.) jezikovno razslojenost prikazuje s socialno, funkcijsko, prenosniško, časovno in mernostno zvrstnostjo. Ta pogled je razviden tudi v drugih slovenskih jezikovnih priročnikih, kot sta Slovenski pravopis (SP) in Slovar slovenskega knjižnega jezika (SSKJ). V sodobnosti se s stilistiko prekrivajo besediloslovne, žanrske in diskurzivne teorije, ki z novih vidikov, predvsem pa z drugimi koncepti in poimenovanji razpoznavajo jezikovno variantnost ter širijo stilistična spoznanja v postrukturalistični, diskurzni stilistiki, ki tej variantnosti iščejo nove metode ter tudi sistematike in tipologije, predvsem v t. i. diskurznih tipih, ki presegajo meje tradicionalne klasifikacije funkcijskih stilov oz. zvrsti (prim. Katnić-Bakaršić 2007: 68).

V postmoderni je tovrstno klasificiranje še posebej težavno, pa tudi problematično (prim. Kalin Golob 2013). Ko na primer Cook (2001: 33, 39) razpravlja o oglasih, jih označuje kot parazitski diskurz, saj si iz drugih žanrov in posameznih besedil sposojajo toliko značilnosti, da jim grozi, da ne bodo imeli svoje stilne identitete. Opozarja, da oznake parazitski ne smemo jemati slabšalno, saj številni postmoderni diskurzi kažejo enake spremembe, gre za brikolaž žanrov, zvrsti, »jezikov«. Zato številnih sodobnih besedil ni več mogoče analizirati in opisovati s pomočjo nekoč veljavnih opozicij: estetsko – utilitarno; umetnostno – neumetnostno; dejstveno – umišljeno, pravilno – napačno ali visoko – nizko (Čmejrková 2000: 26). V tovrstnem raziskovanju ni prostora za spraševanje, kje so meje

2 Tako izhodišče je seveda mogoče le ob predpostavki strukturalnega pomenoslovja, ki leksikalni pomen ločuje na denotativni in konotativni pomen. Pri nas to ločnico postavlja znotraj slovenskega leksikalnega pomenoslovja Vidovič Muha (2000: 45–110), ki za jedro slovarskega pomena določa denotativni pomen: »Denotativni pomen določa dejstvo, da leksem kot jezikovni znak izraža odslikavo (predstavo) spoznavnega objekta, ne da bi spoznavni subjekt (človek) kakorkoli vplival, se pravi po lastni presoji modificiral to odslikavo, npr. čustveno vrednotil, uporabil možnost stilizacije idr.« (ibid.: 46). Medtem ko naj bi s konotativnim pomenom človek po lastni presoji modificiral spoznavni objekt, tako da je za tovrstne lekseme značilna »dvojna pomenska obremenitev: ob načeloma obveznem denotativnem pomenu izražajo še konotatitivni pomen ...« (ibid.: 97). Medtem ko Vidovič Muha adherentno konotativnost razume omejeno v slovarskem smislu in jo pripisuje neprvim pomenom posameznih leksemov, pa je Akhmanova adherentno zaznamovanost razumela širše v smislu sobesedilnega stilnega učinkovanja, torej ne kot razmerje med besedami oz. njihovimi pomeni, ampak kot »odnos med stvarmi«, ki je ustvarjen v ko(n)tekstu (Akhmanova 1976: 46).

sprejemljivosti mešanja jezikovnih sistemov in žanrov, ampak moramo te sisteme na sodobnih besedilih in žanrih najprej poglobljeno raziskati ter pri tem kakršnokoli metodološko utemeljeno klasificiranje in tipologije razumeti le kot pomoč in »solidni temelj za boljše razumevanje stila, diskurza in jezika nasploh« (Katnić-Bakaršić 2007: 71).

V ospredju razmislekov za nove poglede na diskurzivne prakse je opis jezikovne rabe v njih. Na podlagi velike količine podatkov, kar omogoča korpusni pristop, lahko s kombinacijo kvantitativne in kvalitativne metodologije razpoznavamo pogostejše in običajne ter redkejše oz. neobičajne rabe znotraj posameznih diskurzov ter na tej podlagi izdelamo opis, ki identificira te rabe in njihove tendence ter tako daje usmeritev tudi za njihov morebitni leksikografski opis.

Stilistika je v svojem raziskovanju prešla saussurjevsko dihotomijo jezikovni sistem (langue) – jezikovna raba (parole) oz. chomskyjansko delitev na kompetenco in performanco ter prav s stilom nazorno prikazala, da obstajajo pojavi med langue in parole, »ali še natančneje: ki so oboje« (Hoffmanová 1997: 151). Številni avtorji (Michel, Jedlička, Hausenblas, nav. po Hoffmanová 1997: 151–152) to vmesnost razumejo kot sistemske jezikovne pojave, katerih stilna raznolikost izhaja iz jezikovnega sistema (naglasne oblike, oblikoslovne različice ipd.), a so kot stilno opazna kategorija realizirani šele v konkretnem jezikovnem pojavu, v katerem jih lahko analiziramo.[3] Pri tem izgubljajo izpred oči vse »nesistemske« jezikovne pojave v jezikovni rabi, ki jih danes natančneje definirajo k diskurzu usmerjene raziskave. Ali s Koroščevimi besedami (1998: 8), ko definicijo stila postavlja s šestih vidikov: stil je mogoče opazovati le na ravni jezikovne rabe, rezultata jezikovne dejavnosti, tj. v sporočilu, komunikatu.

Pogledi na stil in stilistično raziskovanje se enako kompleksno kažejo tudi v leksikografiji. Deloma na to vplivajo razlike med zahodnim in vzhodnim pojmom stila. Pri prvem je bil dolgo v ospredju stil kot atribut avtorskega, individualnega stila, kar je povezano z usmerjenostjo k raziskovanju leposlovnih besedil, pri drugem je bistvene premike pri teoretskem razmisleku postavljal PLK s funkcijskimi stili kot objektiviziranimi področji jezikovne rabe, torej razumevanje interindividualnega stila, pri nas s Toporišičem poimenovanega kot funkcijska in druge jezikovne zvrsti. Danes se je z novimi metodami in pretokom znanja v jezikoslovju zahodni in vzhodni svet zbližal, stilistika in pogledi na jezikovno variantnost pa lahko veliko pridobijo z dostopnostjo velikega vzorca besedil v elektronskih besedilnih korpusih posameznih jezikov, ki lahko variantnost raziskujejo deloma avtomatizirano, statistično podprto glede na notranjo gradivno členjenost (različne vrste besedil, čas, prenosnik, tema ipd.) ter v kombinaciji s tipično sobesedilno pojavnostjo.

3 V tem smislu gl. npr. za slovensko besedotvorje Logar (2006).

2 STILISTIKA IN LEKSIKOGRAFIJA

Razlikovanje med pogledi, predvsem pa različna leksikografska in jezikovnostilna tradicija ter različna poimenovanja in z njimi koncepti posameznih jezikovnih variant in rab tudi danes odpirajo vprašanja leksikografskemu delu. Treba je namreč pretresti jezikovno raznolikost v dobi, ko klasične, velikokrat jezikovni didaktiki prilagojene modele jezikovnih plasti/zvrsti/registrov vedno bolj zamenjujejo kompleksni pogledi na diskurz in njegovo stilistiko (prim. Hoffmanová 1997: 174–175; Katnić-Bakaršić 2007: 59), govorjenim in pisanim besedilom so se pridružili še hiperteksti, omogočeni z novimi tehnologijami konca 20. in začetka 21. stoletja z vso zapletenostjo hibridnih jezikovnih podsistemov sodobnega časa.

Leksikografski pogledi na jezikovne različice, ki izhajajo iz tipov besedišča, zajetih v slovarsko obravnavo, so odvisni od jezikoslovnih teorij, vplivajočih na slovarsko delo, od vrste slovarja in njegovega uporabnika, njim je treba jezikovno variantnost uslovariti in jo kar se da nazorno predstaviti. Slovarji so za ta namen kot standardne, čeprav ne edine možne, sprejeli oznake oz. kvalifikatorje, ki kažejo »da je dana leksikalna enota stilno označena, strokovna ali časovno opazna« (Žmigrodzki 2010: 12).

Namene in tipe tovrstnega označevanja leksikografski teoretiki in praktiki razumejo različno. Praktični vodič leksikografije (Sterkenburg 2003) razume npr. stilno oznako kot rabno oznako in kazalnik jezikovne enote, ki je »estetsko razlikovalna« (ibid.: 415), rabo pa kot »način, po katerem je leksikografska enota običajno rabljena, da se opomeni časovno, krajevno, stilno, registrsko itd.« (ibid.: 418). Malmgren (2009: 98) stilne in druge rabne oznake definira kot odklone od »običajne« ali neoznačene leksike in za švedske enojezične slovarje navaja, da je oznaka uporabljena, če je beseda ali njen pomen nefrekventna, pogovorna, žaljiva, formalna, starinska ali omejena na regionalne različice: »Stilne in rabne oznake lahko razumemo kot opozorila pri tvorjenju besedil. Zelo pomembno je, da uporabnike opozorimo, da je beseda lahko žaljiva« (ibid.).

Slovar leksikografije (Hartmann in James 1998) ob definiciji stilne oznake (angl. *style label*) kot oznake za stilno raven leksema v slovarju dobro ponazarja zapletenost leksikografije in stilnega označevanja: »Stil je notorično težko opredeliti in slovarji so imeli težave z označevanjem vidikov jezikovne rabe z enotno stilno oznako. (Je 'pogovorno' lastnost stila ali stopnje formalnost ali celo družbenega statusa?).«

Atkins in Rundell (2008: 183) skicirata tipe besedišča, ki jih slovarji vsebujejo, glede na področja, regije, narečja, register, stil, čas, sleng in žargon, odnos, žaljiva poimenovanja. V opisu posameznih tipov in njihovih uresničitev (ibid.: 184–186) je očitno, da večina teh tipov vsebuje stilistične informacije v širšem pomenu besede (leksika glede na izbor formalnega ali neformalnega registra,

nekateri k najmanjši stopnji formalnosti prištevajo tudi žargon in sleng; literarne, birokratske, novinarske in druge stile; časovno opazno leksiko (arhaične, historične, nove); leksiko, ki kaže (ne)naklonjenost do predmeta govora (odnos); žaljive izraze, povezane s politično korektnostjo, tabuje, kletvice ipd.). Avtorja ugotavljata, da pri naboru oznak za posamezne tipe besedišča ni ustaljenih standardov, ni absolutnih vrednosti na lestvici, ampak se o njih odloča vsakokratni leksikografski kolektiv glede na tip slovarja in uporabnike (ibid.: 185, 186).

Klasični tiskani slovarji torej tipe besedišča označujejo z oznakami, ki jih slovarji navajajo in/ali razlagajo v uvodu v slovar ter v slovarju postavljajo običajno pred pomenski del iztočnice. Enako ravnajo tudi nekateri modernejši slovarji. Poljski akademijski slovar je npr. za oznake analiziral prakso v predhodnih poljskih slovarjih in se na tej podlagi »odločil za najboljšo prakso«. Kot pravilo so oznake postavljene pred pomensko definicijo (Żmigrodzki 2010: 19). V sodobni s korpusi podprti leksikografiji pa so se možnosti označevanja spremenile, neomejenost s prostorom na papirju omogoča enostavnejše ponazoritve jezikovne variantnosti, ne le učenih in nerazumljivih okrajšav kvalifikatorjev, predvsem pa je analiziranje dejanske jezikovne rabe postalo objektivizirano zaradi ogromne količine jezikovnih podatkov, ki jih je mogoče s korpusi pridobiti in razčleniti (več o tem Kosem 2015a).

2.1 Oznake v slovenskih slovarjih: od Pleteršnika do druge izdaje SSKJ

Že Pleteršnikov slovar vsebuje 183 »kratic«,[4] ki med številnimi področnimi, slovničnimi in takimi, ki kažejo na jezikovni izvor leksema, navajajo tudi take, ki kažejo stilne posebnosti (figurativno in preneseno, ironično, zaničljivo). Nekatere izmed njih je Šolar uporabil v poskusni redakciji oz. elaboratu za slovar knjižnega jezika (1951, nav. po. Suhadolnik 1997: 562–564) ob delu za slovar slovenskega knjižnega jezika: »V elaboratu so uporabljene slovnične, avtorske, terminološke ipd. oznake (kvalifikatorji), po večini v obliki krajšav /.../ Nekaj vrednostnih in zvrstnih oznak je prevzel iz Pleteršnikovega slovarja (*glavanja* zaničlj., *glavinar* psovka), drugačnih oz. novih ni /.../ Vsega skupaj je v elaboratu nekaj nad 200 različnih oznak, uporabljenih ok. 570-krat.«

Ob poskusnem snopiču in po izidu prve knjige SSKJ (Pogorelec 1964a; Novak 1963; Košmrlj in Müller 1972; Müller 1996; 2009) so bila vprašanja v zvezi z oznakami »osrednje in verjetno najbolj tehtno področje slovenskih kritik v zvezi s slovarjem« (Müller 2009: 18). In čeprav so kvalifikatorji po Müllerju (ibid.) temeljna kvaliteta SSKJ, je »vendar potrebna sistemske in distribucijske preureditve in prevetritve. Kritiki so se ustavljali zlasti ob kvalifikatorjih publicistično, ekspresivno, knjižno, pesniško.«

4 http://bos.zrc-sazu.si/c/PL/seznam_krajsav.html (dostop 4. 8. 2015).

Osnovna težava kvalificiranja, kot jo ob poskusnem snopiču ugotavlja Breda Pogorelec (1964: 237), je premalo enotna in premalo trdna teoretična zasnova slovarja, ki bi tudi glede kvalificiranja morala postaviti jasno točko »od besed, njihovih zvez in pomenov, ki so v standardnem knjižnem jeziku[5] nevtralni, ki tvorijo nevtralno plast knjižnega jezika (osnova slovarja), do besede, ki kakorkoli niso nevtralne: najprej do tistih, ki izražajo določene pomenske odtenke emotivne narave na standardni ravnini (ekspresivno z razvejanostjo pomenov), nato pa do tako imenovanega privzdignjenega jezika v eni smeri, v drugi smeri pa do stilno nižjih plasti (tudi vulgarno).«

Kljub temu, da je SSKJ po kritikah poskusnega snopiča dopolnil kvalifikatorje in razdrobil čustvenostne oznake, je ostalo veliko nerazrešenih mest pri posameznih ekspresivnih (pogosto je bila kritike deležna splošna oznaka *ekspr.*), časovnih in frekvenčnih (nejasnost gradiva za ugotavljanje frekvence, nenatančnost teh oznak glede na počasno izhajanje SSKJ, nejasno ločevanje med *zastarelim* in *starinskim*) in zvrstnih oznakah (*publicistično*, *pisarniško*), ki so brez pravih kriterijev tako označevale leksiko v smislu jezikovne kritike, ne pa stilne umeščenosti (prim. Vidovič Muha 2009: 21; Humar 2009: 66).

Enako prakso nadaljuje tudi SSKJ 2, ki oznak iz SSKJ ne spreminja bistveno, s *publ.* denimo še vedno označuje jezikovno in stilno šibke ubeseditve, torej normativnost, ne pa za poročevalska besedila tipične leksikalne enote, kar bi sodilo v takratni koncept informativno-normativnega slovarja (prim. Kalin Golob 2015). Skupine oznak so v obeh slovarjih ostale enake, posamezne manjše spremembe so le pri preimenovanju stilno-plastnih v stilno-zvrstne kvalifikatorje, ni več frekvenčnih oznak (*raba narašča*, *raba peša*, *redko*), različni so posamezni zgledi in definicija žargonizmov (prim. Ahlin et al. 2014). Bolj kot oznake same na sebi pa ostaja problematično izhodišče označevanja in razumevanje knjižnega jezika v povsem drugačnih jezikovnih okoliščinah sodobnega časa (več o tem v prispevku Gorjanc et al. 2015).

2.2 Predlogi za označevanje jezikovne variantnosti v novem slovarju slovenskega jezika

Ob slovaropisni tradiciji, ki jo je postavil SSKJ ter jo kljub pomanjkljivostim nadaljuje in hkrati ruši SSKJ 2 (prim. Ahlin et al. 2014; Krek 2014), je za nadaljnji razmislek o kvalificiranju leksike treba umestiti tudi predloga, ki sta nastala v

5 Čeprav danes take sintagme v jezikoslovju ne najdemo več, ampak bodisi knjižni bodisi standardni jezik, je raba Pogorelčeve razumljiva v obdobju, ko se je teorija knjižnega jezika pod vplivom češkega jezikoslovja šele uveljavljala ob delu za slovar, medtem ko je v zahodnem jezikoslovju za isti pojem v rabi izraz standardni jezik. V nadaljnjih objavah Pogorelec uporablja poimenovanje knjižni jezik.

prizadevanju za nastanek novega slovarja slovenskega jezika, to sta Predlog za izdelavo Slovarja sodobnega slovenskega jezika (Predlog SSSJ, Krek et al. 2013b), ki je bil prvič javno predstavljen maja 2013, in Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika Inštituta za slovenski jezik Frana Ramovša ZRC SAZU (NSSKJ, Gliha komac et al. 2015), ki je bil predstavljen marca 2015.

Predlog SSSJ navaja (str. 34), da dobi uporabnik prek oznak »osnovno informacijo o tem, v kakšnem kontekstu se beseda običajno rabi in kako učinkuje v realni komunikaciji«. Predvidene so naslednje oznake (str. 94, 95): področne (raba, omejena na določeno strokovno področje), diskurzne ali kontekstualne (omejitev na tipične sporočanjske situacije, besedilne tipe ali diskurze), slovnične, stilne (te so registrske in konotacijske), pragmatične in časovne. Novost predloga je ob spremembi tipov oznak tudi razmislek o smiselnosti klasičnih oznak oz. njihovem nadomeščanju z razlagami v pomenskem opisu besede oz. z vizualizacijo, ki jo omogoča spletna oblika slovarja (str. 34, 93–96).

Jasno je, da Predlog SSSJ izhaja iz odmika od zvrstnega modela, ki pa je v 60. letih ob delu za SSKJ pomenil enako pomemben, tako rekoč revolucionaren odstop od dotedanjega pogleda na jezik, jezikoslovje in znotraj tega jezikovno razslojenost, kot razmislek avtorjev Predloga SSSJ. Če je takrat ničto točko za raziskovanje zaznamovanosti predstavljala nevtralna stilna plast knjižnega (pisnega) jezika, danes iskanje ničte točke ob izhodiščih predloga za t. i. komunikacijsko-informativni slovar (Gantar 2015) ni več tako enostavno. Kot smiselna se kaže misel Marka Stabeja (2015c) v gradivu za eno izmed stilističnih srečanj:

> Morda bi lahko za izhodišče privzeli domnevo, da pojmovanje *knjižnojezikovne* (ali *standardnojezikovne,* na tem mestu je vseeno) *nevtralnosti* zajema predvsem morfonološko in morfosintaktično raven opredeljevanja besedišča. Na ravni besedilne oz. diskurzivne vloge besedišča pa vsesplošna nevtralnost, odvezana sleherne stilistične ali pragmatične vloge tako rekoč ne obstaja. Empirično pa seveda obstaja običajnost (standardnost v statističnem pomenu) oz. neobičajnost rabe leksike v posameznih diskurzih (pa še to je le približek povprečja); to običajnost lahko pretvorimo v usmerjevalno označevanje primernosti oz. neprimernosti določene leksike za določene diskurze. Če to domnevo privzamemo kot izhodišče, moramo imeti za označevanje oz. neoznačevanje leksike a) podatke o diskurzih in b) izdelano njihovo delovno razvrstitev oz. taksonomijo.

V Osnutku koncepta NSSKJ so kvalificiranju besedišča namenjene strani 62 do 70. Predlog ločuje štiri vrste oznak: slovnične oznake, slovnična pojasnila, kvalifikatorje in kvalifikatorska pojasnila (str. 62). Kvalifikatorji in kvalifikatorska pojasnila rešujejo socialno, funkcijskozvrstno in stilno raznovrstnost leksike (str. 64): »Predlagamo torej naslednjo tipologijo kvalificiranja: socialnozvrstni,

funkcijskozvrstni, ekspresivni, časovni in normativni kvalifikatorji ter kvalifikatorska pojasnila.«

Na str. 66 je podrobneje pojasnjeno:

> Leksikalne enote, njihovi pomeni in podpomeni, ki se pojavljajo npr. zlasti v publicističnih, poslovno-uradovalnih ali umetnostnih besedilih, so označeni z opisnim kvalifikatorskim pojasnilom, npr. v poslovno-uradovalnih besedilih, v publicističnih besedilih, v umetnostnih besedilih. Kadar jim je mogoče pripisati točno določene tipske oznake, jih čim bolj natančno opredelimo, npr. v publicističnih besedilih s področja glasbe oz. v reportažah, v poeziji, v oglasnih besedilih ipd. Če se bo pri urejanju oz. redigiranju gradiva izkazalo za relevantno, bomo za tovrstne leksikalne enote, njihove pomene in podpomene uvedli samostojne kvalifikatorje, tj. poslovno-uradovalno (posl.-urad.), publicistično (publ.), umetnostno (umetnostn.).

V Tabeli 1 je prikazana primerjava med različnimi tipi kvalifikatorjev v inštitutskih slovarjih s knjižnojezikovnim izhodiščem.

Tabela 1: Skupine kvalifikatorjev v SSKJ, SSKJ 2 in NSSKJ.

SSKJ 2	SSKJ 2	NSSKJ
slovnični	slovnični	slovnični
pomenski	pomenski	
terminološki	terminološki	znotraj funkcijskozvrstnih
stilno-plastni	stilno-zvrstni	socialno-, funkcijskozv.
ekspresivni	ekspresivni	ekspresivni
časovno-frekvenčni	časovno-frekvenčni	časovni
posebni normativni	posebni normativni	normativni

NSSKJ sicer pri posameznih vrstah kvalifikatorjev želi izboljšati nekatere oznake, npr. *knjižno* v *ozko knjižno* (uporabnik je težko doumel pomen oznake *knjižno* v SSKJ), ukinja *zastarelo* in navaja le *starinsko* (v SSKJ je bila težko določljiva razlika med obema), deloma uvaja tudi kvalifikatorske opise v smislu žanrske (besedilnovrstne) oznake in kombinacijo kvalifikatorja s kvalifikatorsko oznako, kot navaja že Predlog SSSJ. Kvalifikatorja *žargonsko* ni več med socialnozvrstnim oznakami, nadomestila (?) ga je oznaka za *sleng*. Hkrati pa NSSKJ nadaljuje prakso SSKJ in pri socialni zvrstnosti navaja niz oznak, ki se zdijo s stališča uporabnika preveč strokovne, da bi lahko brez težav razumel posamezne variante. Praksa tujejezičnih slovarjev, ki navajajo v okviru registrskih oznak nekaj stopenj formalnosti, običajno vsaj tri (Atkins in Rundell 2008: 185: običajno ena bolj formalna od nezaznamovane, in dve bolj neformalni, npr. *neformalno*, zelo *neformalno*), se zdi

zato s stališča uporabnika boljša. Pri ekspresivnih kvalifikatorjih še vedno ostaja kvalifikator *ironično*, ki je problematičen, saj je ironija po definiciji ustvarjena v kontekstu, zato ni inherentna leksikalni enoti, kot bi se zdelo po načinu kvalificiranja (npr. str. 67: *cvetka*, *idila*, ki ju NSSKJ navaja kot zgled za kvalifikator *ironično* in kvalificira pomen konkretne rabe, ne pa izraza kot takega. Seveda nista ironični sami po sebi, ampak le v kontekstu, iz katerega je jasno, da tvorec misli prav nasprotno od izrečenega oz. zapisanega).

V NSSKJ je tako deloma razvidno odzivanje na nekatere kritike kvalificiranja v SSKJ, večinoma pa je ohranjen koncept jezikovne zvrstnosti, kot ga je postavilo češko jezikoslovje in sprejelo slovensko v 60. letih 20. stoletja. Razumevanje kvalifikatorjev še vedno predpostavlja zelo poučenega uporabnika slovarja, ki si bo vzel čas in najprej preštudiral uvodne razlage o njihovem pomenu (prim. razdelek o slovarskih uporabnikih v tej monografiji). Modernejše pristope želi uvajati s kvalifikatorskimi pojasnili, a so v konceptu vidne zadrege pri njihovi konkretni domišljenosti.

3 OZNAČEVANJE VARIANTNOSTI JEZIKOVNE RABE V SLOVARJU SODOBNEGA SLOVENSKEGA JEZIKA: OD PREDLOGA K REALIZACIJI

Označevanje variantnosti jezikovne rabe, kot ga prikazujemo v nadaljevanju, temelji na zasnovi slovarja in še prej slovarske baze, ki v izhodišče slovarskega opisa za razliko od SSKJ-jeve informativne normativnosti postavlja komunikacijsko oz. sporočanjsko informativnost. Prvi poskus v tej smeri, ki se oddaljuje od knjižnojezikovnega izhodišča kvalificiranja slovenske leksike, je bil izveden že v času izdelave Leksikalne baze za slovenščino (LBS)[6] v okviru projekta Sporazumevanje v slovenskem jeziku, seznam oznak, njihova vrednost in možnosti, ki jih pri prenosu v slovar predstavlja spletna zasnova, pa so bile nato predstavljene tudi v Predlogu za SSSJ (Krek et al. 2013b). Strategije, ki bi leksikografom omogočile konsistentne, jasne in enotne odločitve v zvezi s kvalificiranjem variantnosti jezikovne rabe, v času izdelave LBS niso bile dokončno izdelane, saj je bil temeljni namen LBS na konkretnem gradivu šele preizkusiti opise jezikovne rabe, ki bi izhajali iz analize velike količine besedil in bi temeljili na orodjih, s katerimi je mogoče prepoznavati regularnosti in posebnosti ne samo na ravni skladenjskega vzorčenja, ampak tudi na ravni stilnega in vrednotenjskega potenciala, ki si ga deli jezikovna skupnost kot celota oz. ki prihaja na površje v določenih besedilnih tipih, žanrih in diskurzih. Začenši tako rekoč z ničte točke, je bila ena od temeljnih nalog pri izdelavi LBS tako tudi premislek, kako kvalifikacijo leksike z zgoraj omenjenih vidikov uspešno postaviti v okolje, v katerem se je s SSKJ vzpostavila in zakoreninila knjižnojezikovna zvrstna

6 Več na http://www.slovenscina.eu/projekt (dostop 3. 8. 2015).

kvalifikacija, temelječa na spoznanjih PLK. Ta kvalifikacija je bila kljub nekaterim šibkim točkam in kritikam namreč v SSKJ dosledno izpeljana, hkrati pa je našla svoje mesto tudi v šolskih učbenikih in drugih jezikovnih priročnikih (slovnica, pravopis). Zavedajoč se zakoreninjenosti tradicije zvrstnega označevanja in posledično navajenosti slovarskih uporabnikov na sistem SSKJ-jevih kvalifikatorjev na eni strani in nujne neuspešnosti prenosa knjižnojezikovnega izhodišča na sodobno korpusno gradivo, smo pri izdelavi LBS besede, zveze in pomene označevali glede na različne okoliščine sporočanja, predvsem glede na doseganje stilnega učinka v sporočanjski situaciji ter glede na lastnosti in posebnosti diskurza, zlasti ko je ta omejen na določeno področje, interesno skupino ipd. Poleg tega pa še glede na vrednotenje vsebine in udeležencev v sporočanju ter glede na sporočanjski učinek besede, zveze ali pomena. Za razliko od SSKJ nam je edini jezikovni filter pri vključevanju v slovar predstavljala relevantnost jezikovne rabe (dejstvo, da se beseda/pomen dejansko uporablja), pri opisu in kvalifikaciji pa analiza večjega števila jezikovnih situacij, v katerih se beseda oz. njen pomen uporablja.

Da bi bilo mogoče ugotoviti, v katerih primerih se določena raba uveljavlja in je na ravni jezikovne skupnosti prepoznana kot stilno, časovno, vrednotenjsko ali kako drugače specifična, ter izdelati konsistenten način označevanja, je bilo najprej treba izhodiščne premisleke, načine označevanja ter ubeseditve posameznih oznak preizkusiti na določeni količini besedišča. Ta del je bil, kot rečeno, dejansko opravljen pri izdelavi LBS in načelno povzet v Predlogu za SSSJ, na tem mestu pa želimo osnovna izhodišča dopolniti ter jih nadgraditi s konkretnimi slovarskimi rešitvami. Pri tem se osredotočamo na označevanje kontekstualne in diskurzno specifične ter stilno opazne rabe.

Izhodišče pri označevanju in kvalificiranju leksike nam predstavlja že omenjeni premik od določanja zaznamovanosti v smislu ustrezno – neustrezno k prepoznavanju komunikacijskega učinka pri rabi določene besede oz. njenega pomena v besedilu. Pri določanju načina označevanja in pri umeščanju informacije o rabi v zgradbo geselskega članka smo upoštevali naslednje možnosti:

1. Uvrščanje informacije o jezikovni rabi, v tiskanih slovarjih tradicionalno vezane zgolj na oznako oz. kvalifikator/kvalifikatorsko pojasnilo, v različne segmente geselskega članka, in sicer:
   a) **v oznako** na ravni iztočnice ali pomena, pri čemer so različni tipi oznak v spletni vizualizaciji lahko prikazani na različne načine, npr. z razporeditvijo na zaslonu, z različnimi barvami, fonti ipd., npr.
      - za označevanje tipičnega (pomenskega) konteksta oz. vezanosti na določeno tematiko (t. i. kontekstualne oznake):
        izsušen
        1 ki ima premalo vode; dehidriran

o organizmu
izsušena [sluznica, koža, polt]
izsušene [ustnice]
2 ki nima vode
o vodnih virih
izsušena [struga, puščava, zemlja, njiva]
izsušeno [močvirje, korito, jezero, blato]

- za označevanje vezanosti na tipične sporočanjske situacije (diskurze) ali besedilne tipe (t. i. diskurzne oznake):[7]

bobnati
3 veliko govoriti
v političnem kontekstu
bobnati v [parlamentu, vladi]
lajkati
1 označiti z znakom za dvignjeni palec; všečkati
v družbenih omrežjih

- za označevanje vrednotenja ubesedene predmetnosti (t. i. stilne oz. konotacijske oznake):

dedek
1 oče matere ali očeta
1.1 začetnik
1.2 starec
izraža naklonjenost
Neki stari dedek naju napoti v hotel.
V hišici je živel star dedek z dolgo belo brado.

- za označevanje rabe nekaterih besed s pomensko vlogo poudarjanja, in sicer lastnosti ali količine, vsebovane v besedi, na katero se poudarek nanaša (t. i. konotacijske oznake):

gladko
1 brez težav; z lahkoto
za izražanje poudarka
Trebanjci, ki so doma gladko ugnali Kopрčane.
»Če bi se hotela prijaviti na agenciji za manekenke in modele, bi me gladko zavrnili,« pravi Twiggy.

7 Za ta tip označevanja je značilna tanka meja prehajanja med rabo v določenih prepoznavnih diskurzih in področno oz. terminološko vezanostjo izraza ali njegovega pomena. Zato se zdi ohranjanje pretočnosti med obema tipoma označevanja, ne da bi se leksikograf moral odločati le za en ali drugi tip oznake (tj. med področno in diskurzno), rešitev, ki se približuje realnosti jezikovnih situacij in ne teži k metaslovarski uniformnosti. Pomembneje se zdi na drugi strani ločevanje med načinom ubeseditve dogodkov in stanj, povezanih s posameznim področjem, npr. v medijih, in leksiko, značilno za posamezno področje. V mislih imamo predvsem tipične razaktualizirane lekseme, ki ubesedujejo športne, politične, gospodarske vsebine in sodijo v novinarski diskurz, pa tudi tipične poročevalske avtomatizme (prim. Kalin Golob 2015).

- za označevanje sporočanjskih situacij, ki jih določa družbeni status udeležencev in jezikovne konvencije (t. i. registrske oznake):

tip
3 moški
v sproščeni neformalni komunikaciji
Njej se je zdel tip ful luškan.
Tip je absolutno nor.

- za označevanje rabe besede oz. pomena, ki ga ta dobi zlasti v govorjeni komunikaciji, kamor štejemo tudi izraze v vlogi besedilnih povezovalcev (npr. *mimogrede*), vzklike z besedilno modifikacijsko vrednostjo (*Fant!*), rabo pomena med govorci določene generacijske pripadnosti ipd. (t. i. registrske oznake):

Fant! | Fant moj!
zlasti v govoru
izraža poudarek ali presenečenje
Fant, kako je tečen!
Fant, kako so si nas privoščili.

keš
denar, zlasti gotovina
neformalno; zlasti v komunikaciji mladih
On ma sigurno dost keša, sam ga skriva na računih v tujini zarad davkov
»Našla si bom novo službo za ful keša!«
Ronaldo je edein izmed največjih, pa ne zato ker ma hud auto pa ful keša, ampak zato ker je pač dober.

b) **v indikator**, ki predstavlja poleg razlage samostojno pomensko informacijo, razvidno že na ravni pomenske členitve. Za to možnost se odločamo npr. pri opredeljevanju vezanosti posamezne besede ali pomena na določeno predmetnost ali tematiko (t. i. kontekstalne/diskurzne oznake), ko bi se informacija iz indikatorja v oznaki le ponovila. Tak način omogoča označevanje posameznih pomenov že na ravni pomenskega menija, kar pomeni, da ostaja oznaka, vključena v indikator, uporabniku skrita, ohranja pa se na ravni slovarske baze (gl. spodaj tč. 2), s čimer omogočimo iskanje po različnih tipih oznak tudi v primeru, ko te v slovarju niso neposredno razvidne:

koalicija
1 zveza
1.1 **o političnih strankah**
1.2 **o državah**

<indikator><oznaka tip="**kontekst**">o političnih strankah</oznaka></indikator>
<indikator><oznaka tip="**kontekst**">o državah</oznaka></indikator>

Kvalifikacijo besede ali pomena lahko vključimo v indikator tudi v primerih, ko želimo že na tej ravni opozoriti na pragamatične sestavine pomena (t. i. pragmatične oznake):

fasati jih | Bi jih rad fasal?!
kot grožnja
zelo grobo

<indikator><oznaka tip="pragmatika">**kot grožnja**</oznaka></indikator>

A bi jih rad fasal?« je zdaj zarjovel Semi in to tako, da je Domna in moža odneslo po rumeni opečnati cesti.
Pa kaj ta kmet bi jih rad fasal na gobec!?

c) **v izpostavljeni del razlage**, zlasti npr.

- za označevanje kulturnih, zgodovinskih ali družbeno-političnih okoliščin, znotraj katerih se uporablja določena beseda, njen pomen ali zveza (t. i. kontekstualne oznake):

ikona
2 religiozna slika
v vzhodnoevropski in ruski pravoslavni cerkvi *religiozna podoba ali slika*

Veliki/veliki brat | Veliki brat te opazuje!
po romanu G. Orwella »1984« *izraža popoln nadzor države ali ustanove nad posamezniki*

- za označevanje stilne vrednosti pomena besed in zvez, ki izhaja iz govorčevega vrednotenja ubesedene predmetnosti. Vrsto označenosti lahko v grobem ločimo na táko s pozitivnim (t. i. stilne oznake) in táko z negativnim vrednotenjem, npr.:

potrošno blago
z neodobravanjem *ljudje, dejavnost ali predmeti, ki se obravnavajo kot nekakovostni in hitro pokvarljivi*

lisjak
2 zvit, prebrisan moški
v pozitivnem, hudomušnem smislu
Kako blizu resnice je, da bi trenerski lisjak iz štajerske metropole prevzel vodenje še tretjega prekmurskega prvoligaša?
Naš puščavski lisjak oziroma motorist Miran Stanovnik je pri 46 letih postal očka

č) **v razlago**, zlasti za označevanje pragmatičnih elementov pomena ali frazeološke enote, npr.

izplavati
3 postati opazen, javen
če rečemo, da neka skrita in navadno neprijetna DEJSTVA izplavajo na POVRŠJE, želimo povedati, da postanejo javna in je mogoče o njih kritično razpravljati
Skoraj vsako leto, največkrat v Mladini, izplavajo na površje nove podrobnosti.

Prijateljev primer je brž utonil v policijskih fasciklih nerešenih zadev, Podobnikov pa izplaval na prve strani dnevnega časopisja.

jokati in cviliti

če rečemo, da kdo joka in cvili, želimo povedati, da se pritožuje in tarna, navadno zaradi nepomembnih stvari ali takrat, ko je za ukrepanje že prepozno

Navkljub temu zakonu boste volitve izgubili, in takrat ko boste vi v opoziciji, boste hudičevo na glas jokali in cvilili, kaj smo naredili.
Dokler so vam "dajali" iz njihove sklede "jesti in piti" ste bili tiho, sedaj ko naj bi vas odmaknili od svoje sklede in vam dali čisto svoj krožnik, pa jokate in cvilite.

2. Uvrščanje informacije o jezikovni rabi z upoštevanjem razmerja med slovarsko bazo in slovarjem, ki ustreza razmerju leksikograf – slovarski uporabnik in je pomembno z vidika poenotenja oznak in konsistentnosti ubeseditev na eni in z vidika vzpostavitve različnih iskalnih možnosti na drugi strani. Poleg oznak v indikatorju, ki ostajajo na ravni slovarja skrite, vidne pa so v slovarski bazi (gl. zgoraj primere v b)), sodijo sem tudi oznake, ki so namenjene predvsem leksikografom in notranji ureditvi pomenov. V LBS smo v ta namen uporabljali oznako preneseno, ki napoveduje, da ob izhodiščnem pomenu obstaja še metaforični (tudi metonimični) pomen, ki je z izhodiščnim tako ali drugače asociativno povezan. Vsebuje torej določene pomenske lastnosti, ki omogočajo prevrednotenje pomenskih vsebin na način, da dobijo novo pomensko vrednost, npr.

izvažati
1 prodajati blago v tujino
1.1 prenašati značilnosti česa v druga okolja
preneseno
izvažati [demokracijo, stabilnost]
izvažati [terorizem]

3. Uvrščanje informacije o jezikovni rabi v samostojne zavihke, zlasti v zavihek »Oblika« in »Norma« s podrobnejšo informacijo o standardiziranosti oz. nestandardiziranosti izraza, oblike ali (obliko)skladenjske rabe, ki je posredno tudi informacija o njihovi umeščenosti v kontekstu pisne standardne slovenščine. Primer takega prenosa sta npr. gesli *zajedavec* in *zajedalec*, kjer sta obe možnosti zapisa obravnavani v samostojnih geslih s potencialnimi razlikami na pomenski in kolokacijski ravni, v samostojnem zavihku pa je opozorjeno tudi na (ne)standardnost posamezne oblike.

Prikazani načini kvalificiranja leksike predstavljajo torej način označevanja variantnosti jezikovne rabe, ki se glede na predvideno zgradbo geselskega članka prestavljajo iz izpostavljenih opozoril v klasičnih oznakah tudi na raven drugih

elementov geselske zgradbe. S tem se informacija zlasti pri kontekstualno in besedilnotipsko pogojenih rabah ter znotraj različnih registrov in pragmatičnih elementov pomena zliva s pomenskimi opisi v različnih stopnjah (npr. kot izpostavljeni del razlage ali pa kar kot njen sestavni del) in vključuje v celovito slovarsko informacijo. Tabela 2 zgoraj s konkretnimi primeri prikazani sistem označevanja prikazuje še shematično in predstavlja izhodišče za oblikovanje navodil leksikografom, ki omogočajo pri izdelavi slovarja čim bolj enotne in konsistentne, a hkrati fleksibilne odločitve.

Tabela 2: Označevanje variantnosti jezikovne rabe v predvidenih elementih geselske zgradbe

Vrsta označevanja	**Način in mesto informacije v geselski zgradbi**					
	oznaka	**oznaka v indikatorju**	**izpostavljeni del razlage**	**razlaga**	**Slovarska baza**	**Samostojni zavihek**
Kontekstualno pogojena raba	v krščanstvu, zlasti v grški drami ...	o človeku, o organizmu, o načrtih, o dogajanju in pojavih …	(ikona) **v vzhodnoevropski in ruski pravoslavni cerkvi** *religiozna podoba ali slika;* (kancler) npr. v Nemčiji in Avstriji *predsednik vlade*			
Diskurzivno pogojena raba	zlasti v športu, v političnem kontekstu, v novinarskem žargonu, v družbenih omrežjih …	(sredinec) **v politiki** *pripadnik sredinske neskrajnostne politične opredelitve*				
Besedilnotipsko pogojena raba	v oglasnih besedilih, pogosto v malih oglasih, v kuharskih receptih ...					

Vrsta označevanja	**Način in mesto informacije v geselski zgradbi**					
Konotacijska/ stilna vrednost	za izražanje poudarka, odklonilno, izraža prizadetost, izraža naklonjenost, slabšalno, navadno z neodobravanjem			(anonimen) neizrazit, pust *anonimen predmet ali okolje je neizrazit in brez lastnega značaja ali posebnosti*	preneseno	
Registrsko pogojena raba	v sproščeni neformalni komunikaciji, v formalni komunikaciji in besedilih, zlasti v govoru, zlasti v komunikaciji mladih ...		(vila) **kot odraz premožnosti, statusa** lepa, razkošna hiša	(cviliti) *če rečemo, da glas bilo cvili, želimo povedati, da oddaja neprijeten zvok, navadno zato, ker kdo ne zna igrati*		
Pragmatična sestavina pomena	kot grožnja, grobo, navadno kot zmerljivka …			(ohladitev) *upad navdušenja ali zanimanja, navadno zaradi neprijetnega dogodka*		
Označenost z vidika normativne (ne) ustreznosti						Zavihek Norma, Oblika

ZAKLJUČEK

Proučevanje izraznih sredstev, ki pri nastanku besedila sodelujejo s svojo vrednotenjsko vlogo, je v zgodovini jezikoslovne teorije dolgo in že od nekdaj tudi sestavni del slovarske informacije. Sega od grške retorike prek osamosvojitve v samostojno disciplino – stilistiko – v 19. stoletju do različnih teoretičnih in metodoloških pristopov, zlasti strukturalizma, praškega funkcionalizma, pragmatične stilistike ter funkcijskega modela z usmeritvijo v družbene okoliščine in kontekstualne dejavnike, kot so register, spol in ideologija, vse do t. i. kritične stilistike, usmerjene v opisovanje, manj pa v tipologizacijo diskurznih praks. Prav na opisih komunikacijske vloge jezika, ki temeljijo na poststrukturalističnih modelih, kamor uvrščamo tudi korpusni model, se vzpostavlja težnja po izdelavi slovarskega opisa variantnosti jezikovne rabe, ki presega model vrednotenja – v slovenskem okolju določenega s knjižnojezikovnim izhodiščem – in teži k identifikaciji in opisu dejanske jezikovne rabe ter njenih pomenskih in vrednotenjskih tendenc, ki bi bil predvsem sporočanjsko informativen.

Prvi premik v tej smeri – z novimi metodami, temelječimi na korpusu, in uporabo orodij za njihovo analizo na določeni količini gradiva preizkusiti opis diskurzivnih praks, ki se odražajo v jezikovni rabi – je bil na slovenskem gradivu opravljen pri izdelavi LBS, v prispevku pa smo želeli spoznanja, ki jih je pokazala nadaljnja analiza, prikazati v načinu opisa jezikovne kvalifikacije – v slovarjih tradicionalno vezane na oznake ali kvalifikatorje – v različnih segmentih geselskega članka. Pri tem smo morali upoštevati, da je sistem označevanja variantnosti jezikovne rabe v obstoječih slovarjih za slovenščino (in jezikovnih učbenikih) ustaljen, splošno sprejet in jezikovnim uporabnikom bolj ali manj znan, hkrati pa dejstvo, da modela knjižnojezikovnega izhodišča, ki je temeljil na knjižnojezikovnem filtru že pri določanju knjižnojezikovnosti slovarskega gradiva, ni mogoče prenesti na korpusno gradivo in nove modele komuniciranja v sodobni družbi, na kar so v slovenskem prostoru opozorili prav avtorji prve izdaje SSKJ (Ahlin et al. 2014) in povsem ignorirali avtorji njegove druge izdaje ter snovalci Osnutka za NSSKJ.

V prispevku prikazani model kvalificiranja jezikovne rabe tako ne izhaja več zgolj iz možnosti uporabe različnih tipov oznak, ki so v geselski zgradbi izstopajoča in s tem najbolj eksaktna opozorila, ampak vidi kvalifikacijo jezikovne rabe kot del celovite slovarske informacije, zlasti na ravni pomena – predvsem takrat, ko govorimo o njegovih vrednotenjskih, stilnih in pragmatičnih potencialih. Glede na različnost oznak, ki se uporabljajo v slovarjih, se nam zdi smiselno ločevati terminološke oz. področne oznake, ki lahko razmeroma dosledno sledijo izbrani taksonomiji, in slovnična opozorila, kjer je poenotenje oznak smiselno in mogoče, saj gre za slovnične kategorije, ki povsem na drugačen način določajo rabo konkretnega izraza, kot to velja za konotativne in stilne vsebine pomena. Na

drugi strani bi enotnost ubeseditve na ravni oznak, pa tudi izpostavljenih delov razlage, kadar z njimi opozarjamo na stilno ali vrednostno komponento, pomenila predvidljivost in uniformnost, ki bi sicer dajala vtis poenotenosti, hkrati pa bi leksikografom odvzela fleksibilnost v opisu, ki ga narekuje pogosto težko ulovljiva in razpršena dejanska jezikovna raba.

V izbranem modelu opisa variantnosti jezikovne rabe je treba na koncu izpostaviti še možnosti, ki jih v tem smislu ponuja spletna (mobilna in tablična) zasnova slovarja, ter dejstvo, da je nujna podlaga takemu slovarju obsežna digitalna in interaktivna slovarska baza. Te možnosti se za razliko od klasičnih oznak v tiskanih slovarjih šele odkrivajo in preizkušajo, zato predstavljajo izzive tudi pri oblikovanju spletno zasnovanega slovarja sodobnega slovenskega jezika.

Vrednotenjski pomen in pragmatična funkcija v slovarju

Mojca Šorli

Abstract
In the first section of this paper, arguments are presented in favour of a communicative-pragmatic approach to lexical analysis as it would result in a modern monolingual dictionary with a communicative focus designed specifically for young and/or non-specialist dictionary users. These arguments are based partly on the compilation methods used in the Slovenian Lexical Database as a potential source for new dictionaries of Slovenian, and partly on the Proposal for a New Dictionary of Contemporary Slovenian. In the second section, some conclusions are presented on the functioning of the pragmatic meanings inferred in the doctoral thesis by the author of this article, which discusses at length the (lexicographical) acquisition and interpretation of pragmatic data from the corpus and sheds light on some of the traditionally peripheral aspects of evaluation in language, with a special focus on semantic prosody. The findings are examined in the light of the traditional register and pragmatic labelling in the Dictionary of literary Slovenian and are presented in the context of the Slovenian Lexical Database.

Keywords: corpus analysis, (extended) unit of meaning, pragmatics, connotation, semantic prosody

Ključne besede: korpusna analiza, (razširjena) enota pomena, pragmatika, konotacija, semantična prozodija

1 UVOD

Razmisleki v pričujočem prispevku so vezani na koncept pragmatične funkcije in razširjene enote pomena, kot jo v leksikografskem okviru opredeljujejo J. Sinclair (1996), B. Louw (1993) in drugi (podrobneje o tem v Šorli 2014b: 80–96), ki pa je ostala tako v domačih kot tujih leksikalnih zbirkah in leksikografskih delih, z izjemo sodobnih angleških slovarjev za tujce, večinoma prezrta. Poudariti želimo, da je pragmatika, torej raba, sestavni del pomena, ki ga je zato smiselno v leksikografskem opisu eksplicitno povzeti. V nadaljevanju predstavljamo komunikacijsko-pragmatični pristop k leksikalni analizi, ki bi proizvedel sodobni sporazumevalni tip enojezičnega slovarja, namenjen zlasti mlajšemu in/ali jezikovno nespecializiranemu uporabniku slovenščine, kot smo ga predvideli pri izdelavi Leksikalne baze za slovenščino (2008–2012, dalje LBS).[1] Na rezultatih projekta s številnimi sodelavci (Gantar et al. 2009), ki je bil zasnovan kot konceptualni okvir novega oz. novih slovarjev slovenščine, temelji tudi vsebinski del Predloga za izdelavo slovarja sodobnega slovenskega jezika (Krek et al. 2013b). Pragmatični vidik leksikalnega pomena je bil do sedaj v slovenski leksikologiji in leksikografiji obroben ali le delno obravnavan, zato izpostavljamo tiste postopke leksikalne analize, ki pomenijo odločen premik k večji pragmatični in sporazumevalni vlogi slovarja in za katere menimo, da bi jih bilo treba sistematično upoštevati tudi pri izdelavi novega slovarja slovenščine. Po uvodnem delu povzemamo prakse stilno-pragmatičnega označevanja v LBS (2.1) v povezavi s pomenskimi opisi, ki so zasnovani tako, da omogočajo konsistenten in sporazumevalno usmerjen opis tudi s stališča vrednotenja. V drugem in hkrati jedrnem delu (2.2) predstavimo ugotovitve doktorske disertacije, v kateri smo poglobili pragmatične analize z nastavki pri oblikovanju pomenskih opisov v LBS, tako da smo izčrpno obravnavali pridobivanje in interpretacijo pragmatičnih podatkov iz korpusa in osvetlili nekatere tudi v tujem okolju tradicionalno obrobne vidike jezikovnega vrednotenja v leksikografskem okviru na primeru semantične prozodije (Šorli 2014b), izsledke pa primerjamo z obstoječim stilno-pragmatičnim označevanjem v SSKJ in LBS. Glede na do sedaj prevladujoči jezikovnosistemski pristop v slovenski leksikografiji in sledeč sodobnim, korpusnim spoznanjem o funkcioniranju pomena (npr. Sinclair 1996, 1997; Stubbs 1995, 1995b; Partington 1998; Hunston 2007; Atkins in Rundell 2008; Philip 2009) temelji prispevek v celoti na prepričanju, da potrebujemo uporabniki slovenščine danes nov slovar z opisom jezika s stališča njegove sporazumevalne vloge. Nujne so izboljšave pri vsebini in obliki pomenskih opisov nekaterih povsem vsakdanjih besed, pogosto na ravni jezikovne rabe, kar kažejo tudi nekatere izmed tistih leksikalnih enot (dalje LE)[2]

1 Ena izmed aktivnosti projekta Sporazumevanje v slovenskem jeziku (2008–2013) Ministrstva za šolstvo in šport RS in ESS: www.slovenscina.eu (dostop 23. 7. 2015).

2 V nadaljevanju uporabljamo leksikalno enoto (LE) za označevanje posameznih pomenov, tudi pomena/pomenov večbesednih enot, kot so frazeološke enote (FE), vključno s stalnimi besednimi zvezami (SBZ). LE ločujemo od jezikovne enote oz. leksikalnega niza.

v SSKJ, ki namesto specifičnih konotativnih oznak nosijo oznako *ekspr.* zato, ker »odnos govorca do poimenovanega ni natančno določljiv, npr. *čenča*, *fantastičen* ('ki se pojavlja v visoki stopnji'; 'nenavadno, izredno lep'), *grandiozen*« v Osnutku koncepta novega razlagalnega slovarja slovenskega knjižnega jezika (Gliha Komac et al. 2015: 67; dalje NSSKJ). Za aktualnost slovarja je zelo pomembno, da prepoznamo rabo, ki je tipična danes, ta pa je, kot kaže korpus, prav nasprotna temu, kar npr. pri *grandiozen* kot prvi pomen navaja SSKJ. Prevladuje namreč raba, ki izraža negativno držo govorca do te vsebine (nezaželena velikopoteznost). Pri izdelavi slovarske baze ali slovarja je pomembno, da ugotovimo, kakšno vlogo igra pragmatika pri posameznem pomenu, da bi tako zagotovili ustrezen semantično-pragmatični opis. Pragmatični podatki so integralni del razširjene enote pomena.

1.1 Aksiološki vidiki pomena

Pojem vrednotenja in aksiološki vidiki pomena so ključni za razumevanje, kako jezik pravzaprav deluje. Vrednotenje razumemo na tem mestu v najširšem pomenu besede, kot nekaj, kar celostno zajema govorčev odnos oz. emotivno držo do povedanega, ne zgolj v shematiziranih binarnih opozicijah tipa pozitivno/negativno, slabo/dobro, ugodno/neugodno itd., ki so značilne zlasti za konotativni pomen, temveč tudi širše, v tipično bolj razpršenih in kontekstualno pogojenih oblikah vrednotenja, ki se izražajo v specifičnih, a konvencionalno kodiranih razmerjih med obliko in pomenom. Gre za kolokacijski pomen, ki ga ni mogoče ugotavljati z introspekcijo (Louw 2000). Kot ugotavljata že Kay in Fillmore (1999), je pragmatična moč v veliki meri konvencionalizirana v jeziku.

V doktorski disertaciji smo zato postavili pod vprašaj domnevo, da pragmatike v enojezičnem slovarju ni potrebno eksplicitno opisovati, češ da govorci materni jezik obvladamo intuitivno. To, kar počnemo s pomenom v spontanem jezikovnem sporazumevanju, se bistveno razlikuje od (metajezikovnega) opisa pomenov, tipično v slovarju (Atkins in Rundell 2008: 263). Mogoče je sklepati, da zgoraj omenjena domneva izhaja iz razumevanja pragmatičnega pomena kot dodanega oz. sekundarnega, kar bi pomenilo, da je njegovo navajanje opcijsko. Toda v poglavjih, ki sledijo, razumemo ta pomen, v ožjem smislu vrednotenje, kot sestavni del pomena, znotraj katerega lahko za potrebe pomenske analize sicer ločujemo med semantičnimi in pragmatičnimi vidiki, vendar gre v dejanski komunikaciji za kontinuum oz. tesen preplet obojega. Če je bilo še nedavno razširjeno prepričanje, da je tvorjenje pomena v občutno večji meri kot na primer uporaba pravil v skladnji in fonologiji rezultat zavestnih procesov, pa korpusna analiza nekaterih pragmatičnih pojavov kaže, da ne drži povsem, da tvorci besedil vedno vedo, *zakaj* to, kar so izrekli ali zapisali, pomeni to, kar pomeni. Vprašanje, zakaj potrebujemo pragmatiko v slovarjih, je zato tesno povezano z vprašanjem, zakaj

sploh potrebujemo (enojezične) slovarje in kako jih uporabljamo. Tradicionalno se izpostavljata predvsem kognitivna in sporazumevalna vloga slovarja, pri čemer prva zadovoljuje potrebe po znanju, druga pa omogoča besedilno razumevanje, produkcijo in prevajanje, pa tudi pomen zunajleksikografskih potreb uporabnika, kot jih definirata Bergenholtz in Tarp (2003).[3]

1.2 Slovar kot družbeni artefakt in konvencionalnost pomena

S tem ko v okviru sociolingvističnih teorij raziskujemo jezik kot (družbeni) potencial (Halliday 1978) in ne kot jezikovno zmožnost posameznika ali abstraktni mentalni univerzalizem, je edino smiselno, da dobi pragmatika osrednje mesto tudi v leksikografskem opisu jezika, prej pretežno zaznamovanem s pojmom nujnih in zadostnih pogojev, pri čemer je med pragmatiko, sociolingvistiko, diskurzivno ter konverzacijsko analizo včasih težko začrtati jasne meje. Za slovar kot družbeni artefakt je bistvenega pomena družbeni oz. konvencionalni vidik jezikovne rabe (Hanks 2013). Šele s korpusno analizo je postalo jasno, da pomene v veliki meri razumevamo, ker so produkt družbene konvencije (npr. Levinson 1983; Stubbs 1995c, 2001; Sinclair 1997; Channell 1999; Louw 2000; Philip 2009 itd.), in ne zato, ker bi bili razvidni iz semantične in skladenjske zgradbe pomenskih enot oz. razmerij med njihovimi sestavinami (gl. tudi Šorli 2014b: 105–114). Ilokucijsko moč prepoznavamo, ker je izražena v konvencionalni jezikovni obliki, kar velja tudi za sintagmatske vzorce. Če konvencije sprva ožje namigujejo zlasti na družbeni vidik v kontekstu teorije govornih dejanj, na ravni njihove intencionalnosti in njihovega razumevanja, pa danes govorimo o konvencionalnosti tudi na ravni jezikovnih sredstev, kjer potekajo (nezavedni) pomenotvorni postopki, ki jih je mogoče odkriti šele z vpogledom v veliko število primerov rabe.

2 PRAGMATIČNI VIDIKI LEKSIKALNEGA POMENA

V leksikografskem kontekstu sta mogoča dva osnovna pristopa k stilno-pragmatičnemu označevanju, ki zadeva vrednotenje, tj. kvalifikatorske oznake in pojasnila ali eksplicitno navajanje pragmatičnih vidikov pomena v pomenskem opisu. V okviru prvega so bile v LBS uporabljene naslednje kategorije oznak: diskurzivne oz. kontekstualne in stilne oznake, kamor sodijo tudi registrske oznake in oznake besedilnih tipov, ter skupina pragmatičnih kvalifikatorjev. V tem razdelku najprej

3 Osnovna delitev po teoriji leksikografskih funkcij je na sporazumevalne oz. komunikacijske in kognitivne slovarje.

na kratko predstavimo razlike v rabi pragmatičnih kvalifikatorjev, ki določajo konotacijo LE, v LBS in SSKJ (2.1), v nadaljevanju (2.2) pa se podrobneje posvetimo pragmatiki v pomenskih opisih in sicer tako, da izsledke doktorske disertacije (Šorli 2014b) primerjamo s prakso stilno-pragmatičnega označevanja v LBS (Gantar et al. 2009, 2011; Šorli 2013; Gantar in Kosem 2013) in SSKJ (tudi Uvod XVIII–XXII).

Pragmatično označevanje omogoča prikaz vrednotenjskega pomena na več ravneh. Segment oznak v LBS je namenjen označevanju t. i. sekundarnih pomenov, ki so vezani na besedo ali izraz, a niso nujno v enaki meri skupni vsem uporabnikom v jezikovni skupnosti, tipično konotacija. V nasprotju s temi so intuiciji tipično nedostopni periferni pomeni, ki jih določata kontekst in raba (Philip 2009) in jih predstavljamo v 2.2.

2.1 Konotativne oznake v LBS in SSKJ: stilna zaznamovanost ali pragmatika leksikalnega pomena?

Pri konotaciji gre za delno subjektivne, psihološko pogojene pomene, zato je nabor možnih oznak v LBS odprtega tipa:[4]

MIŠKA
2.1 ženska, dekle ali hčerka
O: ljubkovalno
[mala] miška; miška [moja]

SSKJ ima tu oznako **ekspresivno**: 2 ekspr. *ženska, navadno mlajša, prikupna, ali otrok*, torej brez podrobnejše opredelitve zaznamovanosti.

V nekaterih primerih gre za pragmatične prvine, ki jih je težko ali celo nemogoče opredeliti metajezikovno, s kvalifikatorjem, ker gre za integralni del pomena iztočnice:

CRKNITI
2.2 [umreti]
O: zaznamovano
1 če rečemo, da ČLOVEK **crkne**, na zelo grob način povemo, da umre, da nam je za to vseeno ali da mu to privoščimo[5]

4 Gesla v obliki, navedeni v tem prispevku, temeljijo na uradni končni verziji v formatu xml kot rezultatu aktivnosti LBS in Navodilih za avtorje (Gantar et al. 2011).

5 Vse oštevilčene razlage oz. t. i. pomenske sheme v tem poglavju so, če ni drugače navedeno, povzete po LBS. Gre za predslovarske opise, ki še niso oblikovani v skladu z namenom in funkcijo konkretnega slovarja.

Mojega očeta mi je bilo žal, ne rečem, moja stara pa lahko **crkne** *na licu mesta.*

Povem vam, da bi me možakar tam zunaj pustil **crkniti** *od mraza in lakote, če mu ne bi dal denarja, ki ga je zahteval od mene.*

SSKJ označuje ta pomen (umreti) z oznako **nizko** (za ljudi) in **pogovorno** (za živali), pri čemer se izgubljajo podatki o okoliščinah in s tem o aktiviranem pragmatičnem pomenu.

Za razliko od SSKJ izpostavlja LBS evfemističnost pomenske enote v razlagi:

> **NEČEDNOST** [*navadno množina*]
> [nemoralno, kaznivo dejanje]
> 2 nečednost je zelo mil izraz za kaznivo dejanje ali dejanje, ki izraža človekovo moralno izprijenost

Nekdanji nagrajenec GZS zdrsnil v brezno dolgov in poslovnih ***nečednosti****.*

V Mariboru naj bi razkrili domnevne ***nečednosti*** *nekaterih bivših in sedanjih policistov.*

Policisti trdijo, da so ju že večkrat ovadili zaradi podobnih ***nečednosti****.*

SSKJ izpostavlja zgolj splošno konotativnost oz. konotativno ekspresivnost, besedi pa pripiše tudi oznako **knjižno**:

> SSKJ: knjiž., **ekspr.** ničvredno, malovredno dejanje: počenjati nečednosti; marsikatere nečednosti je kriv alkohol / govori nečednosti // lastnost ničvrednega, malovrednega človeka: njegova nečednost presega že vse meje

Večinoma sodobna publicistična besedila oznake *knjiž.* ne potrjujejo, saj ima na primer Gigafida razmeroma veliko (1.671) zadetkov.

Spodaj lahko ugotovimo, da je *razmišljati* v 4 zgolj slogovno nevtralna nadpomenka od *tuhtati*, ki pa ima svoje pomenske specifike, zato bi bila poleg navedbe udeleženca PROBLEM na mestu tudi navedba okoliščin, v tem primeru načina, na katerega se takšno »razmišljanje« dogaja:

> 3 če ČLOVEK tuhta ali razmišlja o tem, kako rešiti konkreten PROBLEM
>
> **PREDLOG:**
> 4 če ČLOVEK tuhta, zavzeto in navadno dlje časa razmišlja o tem, kako rešiti konkreten PROBLEM

*Od takrat **tuhta**, da bi sestavil vozilo, ki bi prineslo nov rekord in prebilo zvočni zid, saj naj bi doseglo hitrost 1.367,9 km / h.*

*Mož je **tuhtal** cel dan, a do rešitve problema ni prišel.*

*O suknjiču je kar nekaj časa **tuhtal** in se odločal.*

*Umaknil sem se nazaj v temnino za vogalom in sem krčevito **tuhtal**, kaj to lahko pomeni.*

*Začel sem vročično **tuhtati**, kako bi se neopazno znebil onih treh.*

*Njegovi možgani pa ves čas **tuhtajo**, celo ponoči.*

SSKJ ima oznako *pogovorno* ob sinonimni razlagi *premišljati, razglabljati*, pri tem pa vsaj dva od štirih zgledov nakazujeta časovne in načinovne okoliščine, izpostavljene v opisu 4.

Z zgornjimi primerjavami je povezano vprašanje uporabe konotativnih oznak v SSKJ, ki tako konotacijo kot pomene, ki nastajajo v širšem besedilnem okolju, nakazuje z oznako, najpogosteje z *ekspresivno* (gl. tudi *vrenje* 2,[6] *kolovratiti, vrtoglav, sprožiti* 3 in 4, *stokati* 2, *vohati* 2 itd.) ali ponekod s posebnimi ekspresivnimi kvalifikatorji oz. pomensko natančnejšimi oznakami, npr. *slabšalno*, *nizko*, *šaljivo* (*crkniti* 1, *boljša polovica* itd.), drugje celo z oznako *knjiž.* (npr. *sterilen* 2). Ob tem na podlagi analize nabora LE iz LBS in nekaterih drugih ugotavljamo, da pri velikem delu zlasti frazeoloških enot v SSKJ, ki nosijo oznako 'ekspr.', najverjetneje ne gre za konotacijo, temveč semantično prozodijo (gl. 2.2), torej pragmatično funkcijo LE v sobesedilu oz. razširjeni enoti pomena. Domnevamo, da je tako, ker SSKJ ne ločuje med obema ravnema pragmatičnega pomena, nadaljevanje tovrstne prakse pa obeta tudi NSSKJ. Da naj bi ekspresivni kvalifikatorji iz SSKJ načeloma označevali le konotacijo, kot smo jo ožje opredelili v tem prispevku, potrjuje definicija iz koncepta novega slovarja: »Ekspresivna leksika poleg denotativnega pomena izraža tudi subjektivni vrednotenjski odnos govorca do poimenovanega. Ekspresivnost je bipolarna in se vselej giblje na osi pozitivno–negativno« (Gliha Komac et al. 2015: 66). Rabo oz. pragmatični pomen je z enojezičnega vidika upravičeno in smiselno opredeljevati s stilno-pragmatičnimi kvalifikatorji kvečjemu tam, kjer gre za (ekspresivno ali časovno) konotacijo, ki se veže predvsem na pomene, uskladiščene v mentalnem leksikonu govorca. Za SSKJ je značilno, da ne razlikuje med npr. slabšalnostjo kot konotacijo, ter, kot bomo videli v nadaljevanju, (negativnimi) okoliščinami pomena, ki tipično nastopajo pri posameznih LE in imajo ustrezno pragmatično funkcijo (npr. *skotiti* 3: vojna je skotila vse te grozote). Tovrstne vidike pomena bi bilo treba opisovati predvsem v razlagi (v LBS npr.: *enačiti, čas kislih kumaric, (to ni) mačji kašelj, vrtoglav* 1, *sprožiti* 1 itd.), saj gre praviloma za kompleksen preplet semantike, pragmatike in stilistike.

6 Številka ob iztočnici pomeni številko (pod)pomena v citiranem viru.

2.2 Semantična prozodija: vrednotenje in (pragmatična) funkcija

Vrednotenjski pomen se v besedilih tipično izraža tudi v obliki semantične prozodije oz. pragmatične (diskurzivne) funkcije. Poenostavljeno gre za to, da je mogoče tudi navidez nevtralne besede oz. pomene razumeti skozi prizmo vrednotenja ter s pozitivnimi in negativnimi asociacijami v najširšem smislu govorčevega odnosa oz. emotivne drže, ki jih besede pridobijo ob pogostem sopojavljanju z drugimi besedami.

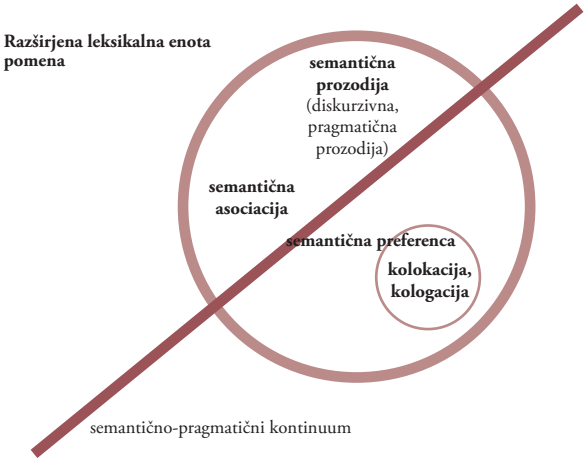
2.2.1 Teoretska ozadja

Semantična prozodija izhaja iz jezikoslovno-leksikografske tradicije J. McH. Sinclairja, temelječe na analizi elektronskih korpusnih besedil, ki ima začetke v delu J. R. Firtha. Iz spletnega statističnega prikaza rabe termina »semantic prosody« skozi čas je razvidno, da raba sovpada z začetkom intenzivnega razvoja korpusnega jezikoslovja in njegovih vplivov na sodobno leksikografijo. Kritičen pregled študij, pogledov, pristopov in razvoja pojma semantična prozodija, ki ga je v prispevku iz leta 1993 prvi predstavil B. Louw, podaja D. Stewart (2010). Tudi ena najpogosteje navajanih definicij prihaja iz tega teksta (1993: 157), namreč »stalen pomenski sij, ki ga kolokatorji ustvarjajo okrog posamezne oblike«. Louw sam se je zavedal, da bodo tako mnogi semantično prozodijo razumeli kot obliko konotacije, v polariziranih kategorijah pozitivnega in negativnega (gl. Stubbs 1995b; Partington 2004; Dilts in Newman 2006 itd.). S poudarkom na njeni pragmatični funkciji pa opisuje Sinclair semantično prozodijo takole:

> Semantična prozodija (Louw 1993: 157) izraža odnos ali držo govorca do izrečenega in je na pragmatični strani semantično-pragmatičnega kontinuuma. V besedilu se realizira na najrazličnejše načine, saj v pragmatičnih izrazih običajne semantične vrednosti besed niso nujno relevantne. Toda ko jo opazimo med raznolikimi načini izraza, je takoj jasno, da ima semantična prozodija vodilno vlogo pri vključevanju jezikovne enote v njeno okolje. Izraža nekaj, kar je blizu 'funkciji' enote – pokaže, kako je treba razumeti njeno funkcijo. Brez nje besedni niz zgolj 'pomeni' – ni uporabljen v realni sporazumevalni situaciji. /.../ Ko smo prišli do semantične prozodije, smo najverjetneje prispeli do same meje leksikalne enote (Sinclair 1996: 13).

Louw (2000: 48) poda svojo definicijo: »/s/emantična prozodija je kolokacijski pojav, in sicer takšne vrste, ki ga je mogoče identificirati z računalniškimi metodami v jezikovnih korpusih, ne z intuicijo«. Vloga prozodije je v tem, da

poveže leksikalno izražene pomene s kontekstom situacije tako, da postanejo neločljivi od konkretne ubeseditve in da se razširijo po celotni enoti (razširjenega) pomena, zaradi česar je Sinclair pojav pravzaprav sploh poimenoval »prozodija« (Sinclair 2004: 117). Izvorno štiridelno razširjeno enoto pomena, ki vsebuje kolokacije, koligacije, semantično preferenco in prozodijo (Sinclair 1996) – od katerih so vse razen prozodije opcijske (Sinclair 1999, 2004) – je Philip (2009) dopolnila še s semantično asociacijo (po Hoeyu *semantic association* (2005: 16)). Čeprav sta semantična preferenca, torej abstrahirana pomenska pripadnost kolokatorjev, s katerimi se beseda oz. LE v besedilih najraje druži, in semantična prozodija ponekod dokaj prekrivni in je bil odnos med obema pogosto opredeljen hierarhično kot razmerje med skupino (preferenca) in podskupino (prozodija), je načelna razlika v tem, da gre pri semantični prozodiji običajno za nek konkreten »scenarij«, torej potek dogodkov ali soodvisnost stanj (Šorli 2012: 103, po Philip 2009). Philip ugotavlja, da imamo opraviti z dvema ravnema pomenske analize, pri čemer je preferenca besedno, prozodija pa funkcijsko, frazeološko in kontekstualno pogojena. To, kar določa semantično prozodijo, je pravzaprav kumulativni učinek vseh ostalih komponent pomena, namreč funkcija pomena v sporazumevalni situaciji:



Slika 1: Razširjena enota pomena (Philip 2009, dopolnjeno po Sinclair 1991, 1996).

Primer: ENAČITI (KAJ S ČIM/KOGA S KOM)
Ob pregledu 300 od skupno 3.138 konkordanc za **enačiti** (4,2 na milijon) v korpusu FidaPLUS izstopajo kolokacije *enačiti* [*sionizem, boj, politiko, državo, človeka, ljudi*], v predložnih zvezah pa *enačiti z* [*državo, vlogo, pojmom, interesom, rasizmom, terorizmom, politiko, ljubeznijo, močjo, položajem, obliko* itd.]. Razen rasizma in terorizma v nizih ne nastopajo kolokatorji, ki bi jih lahko vnaprej

negativno opredelili. V vlogi osebka najdemo [*Freudizem, psihoanaliza, mnogi, ljudje, večina, otrok, šola, država*]:

ična je tista država, ki nastopa na tak način in ki lojalnosti do države ne	enači	z nacionalno pripadnostjo. Sicer pa moj življenjski krog ni le Koper, to so tudi Ljubl
a koprenka. OPOZORILO Je zelo okusna goba in ponekod jo po kakovosti	enačijo	z jurčki. Ker pa je podobna nekaterim strupenim in celo smrtno nevarnim vrstam, na
hnilo t. i. = venetsko teorijo = prepričanje, da lahko prednike Slovencev	enačimo	z antičnimi Veneti. Takšna prepričanja, ki jih avtorji ne morejo podpreti z izsledki p
i) predvsem šport za resne, zanesljive in delavne ljudi (marsikdo ga celo	enači	z garanjem). Nič čudnega, da številni aspiranti in ambiciozni bodoči poslovneži zavz
kalorij odvisen od panoge, s katero se športnik ukvarja; tako ne moremo	enačiti	teka na sto metrov in dviganja uteži. Ogljikovi hidrati so najdragocenejše gorivo za
mnogih primerih pa je pojem nadlegovanja tako zelo spolitiziran, da ga	enačijo	le z družinskim nasiljem in nasiljem nad ženskami. Med tistimi, ki jim zakon in medij
emo, ali se ti reševalci problema beguncev sploh zavedajo, da ni mogoče	enačiti	takratnih slovenskih beguncev in današnjih beguncev, ki jih sprejemamo v razna spr
dstotka BDP. Finančnega učinka sedanjega prometnega davka ne moremo	enačiti	s finančnim učinkom davka na dodano vrednost. Prometni davek je enofazni davek,
za mini krilcih, toda me vztrajamo na svojem bregu. Tako kot ne moremo	enačiti	oziroma metati v isti koš vseh pop izvajalcev ali vseh rock zasedb, tudi ni na mestu
co slovenskega jezika, napisano v latinščini, saj njen avtor v predgovoru	enači	Slovence, Slovane, Venete. Ta slovnica poleg Trubarjevih del in prevoda biblije pom
her poudaril, da Palestina ni primerijiva z Afganistanom in da ju Bush ne	enači	. Ervin Hladnik Milharčič Sklepni dokument plenarnega zbora cerkve na Slovenskem
zpodbijane akte državnega zbora. Ustavno sodišče pojasnjuje, da ne gre	enačiti	postopkov pred ustavnim sodiščem v letih 1994 in 1998 s tokratnim. V prejšnjih post
en dostop. Pri zakonu o vodah se ponovno kaže dejstvo, da morja ne gre	enačiti	z notranjimi celinskimi vodami. Precejšen del naše stroke se zavzema za poseben za
grajo s človeškimi usodami in zavlačujejo, ščitijo ženske, ki bi jih morali	enačiti	z kriminalci. Podatki kažejo, da so zapori 120-140 - odstotno prenapolnjeni in da je
so z nami in v soboto bo z nami zanimiva gostja, ki bi jo lahko prav tako	enačili	z vročinskim valom, saj marsikomu postane vroče, ko jo zagleda ... Slovenska pevka
pozna postseptembrsko bombastičnost oblikovalcev javnega mnenja, ki	enačijo	islam z Al Kaido in muslimane z Bin Ladnom. Gospa Rachel iz Kanade je premožna, p
publikanske armade najtrši oreh mirovnega procesa, ker jo republikanci	enačijo	z vdajo in izdajo glavnega cilja teroristične vojne, ki so jo sprožili pred tremi deset
, zapriseženi sovražnik mutantov, torej moderni Adolf Hitler, ki mutante	enači	z Židi in jim pripravi posebna koncentracijska taborišča. Možje X 2 so torej le znans
anja TOM. Do sedaj smo spremljali rast drobno prodajnih cen, ki smo jih	enačili	z inflacijo oziroma smo ji rekli = R =. Ta je bil osnovni podatek, na osnovi katerega
iso isto kot možnosti. Enakost slednjih so predstavniki gimnazije Poljane	enačili	z enakimi zakoni v članicah Unije, k poenotenju katerih si namreč Unija prizadeva.

Slika 2: Konkordančni vzorec za lemo *enačiti* (referenčni korpus FidaPLUS v programskem okolju SkE).

Podrobnejši pregled okolice potrjuje, da gre pretežno za kontekste, v katerih je mogoče prepoznati negativne okoliščine pomena. Prevladuje vzorec 'kdo enači kaj s čim', iz situacij pomena pa je razvidno, da je s stališča tvorca besedila oz. glede na splošno sprejete norme 'enačenje' vsaj sporno, če ne zgrešeno:

*Arabske države so med pripravljanjem osnutka sklepne deklaracije odstopile od zahteve, da bi sionizem **enačili** z rasizmom.*

*V zahodnih državah muslimansko vero **enačijo** s terorizmom, vse Arabce pa obravnavajo kot potencialne osumljence, ki morajo dokazati svojo nedolžnost.*

*Njen boj zoper režim je bil boj za lastne interese, v prvi vrsti stanovske interese duhovščine, ki jih ta preprosto **enači** z interesi cerkve.*

*Ker dobro spolnost **enači** z ljubeznijo, ohranja odnos s tem partnerjem, pa čeprav je vse v tem odnosu slabo, razen spolnosti.*

Pojavlja se tudi vzorec 'kdo enači kaj (dv. ali mn.)', kot kaže spodnji primer, iz katerega je neposredno razviden prototipičen pomenski scenarij, po katerem nekdo vidi, dojema ali razume dve sicer različni stvari kot eno, čeprav za to ni objektivne osnove:

*Po Jonesovem mnenju načelo ugodja omogoča, da otrok **enači** dve precej različni stvari, ki sta si podobni zato, ker otrok do njiju čuti enako ugodje oziroma interes.*

V LBS smo tako na podlagi navedenih korpusnih podatkov oblikovali opis:

5 če ČLOVEK **enači** KAJ s POJAVOM, POJMOM ali LASTNOSTJO, meni, da gre za enaki stvari, pri tem pa navadno spregleda razlike, bodisi zaradi nevednosti ali namerno, zaradi predsodkov

SSKJ ob tem geslu navaja kratko, sinonimno razlago in zglede, ki z izjemo enega potrjujejo negativno vrednotenje *enačenja*:

imeti, šteti za enako (nepravilno je **enačiti** normo z akordom); po določeni lastnosti imeti kako stvar za enako z drugo (le kako bi **enačili** stroj in človeka// opise v povesti lahko **enačimo** s pisateljevim otroškim svetom).

Raziskovalci semantični prozodiji večinoma pripisujejo vrednotenjsko funkcijo in implicitnost ter poudarjajo, da jo je treba videti zlasti kot verjetnostni pojav, ne kot sistemsko kategorijo v smislu bodisi njene prisotnosti ali odsotnosti (po Flowerdew 2011: 21). Hunston (2007) meni, da je semantično prozodijo najbolje razumeti prav v pomenu diskurzivne funkcije, ki ji ga pripisuje Sinclair, sicer pa kot »enoto pomena, nekaj, kar se upira natančnemu ubesedenju in česar zelo verjetno ne moremo definirati zgolj kot bodisi 'pozitivno' ali 'negativno'«. Če povzamemo po Philip (2009: 3), s posamezno LE povezana prozodija izraža govorčev ali piščev osebni odnos oziroma vrednotenjsko in emotivno držo glede posameznega konteksta ali scenarija, prav tako tudi njegovega (pričakovanega ali dejanskega) rezultata.

2.2.2 Konotacija vs. semantična prozodija

Kot rečeno, naj bi bila konotacija locirana v mentalnem leksikonu govorca in ne konvencionalizirana na način kot semantična prozodija, ki je odvisna predvsem od razmerja med enoto in njenim tipičnim leksikalnim okoljem, medtem ko se konotacija navezuje predvsem na razmerja med besedo (manj pogosto LE) in govorcem/poslušalcem. Seveda prihaja do delnega prekrivanja, saj sta vrednotenjski vidik in stališče govorca pomembna tako za konotacijo kot prozodijo (Stewart 2010: 28–29), toda tu privzemamo prepričanje, da prozodija nikakor ni zgolj vrsta konotativnega pomena. Menimo, da bi se morala razlika med obema tipoma pragmatičnega pomena ustrezno odražati v leksikografskem opisu, kar ponazarjamo s pomensko obravnavo enote *boljša polovica*, v kateri so razvidne razsežnosti potencialnih razlik med konotacijo in semantično prozodijo:

V LBS zveze ni med iztočnicami, zato smo jo naknadno analizirali v korpusih (FidaPLUS 200; 0,3 na milijon; GF 599: 0,96 na milijon), SSKJ pa ponuja:

šalj. to je moja **boljša polovica** - žena; šalj. svojo **slabšo polovico** boš pa že pregovorila – moža

V SSKJ torej dobimo podatek o semantičnem pomenu in konotaciji besedne zveze, ne pa tudi o tem, kaj želi govorec z rabo doseči, torej kakšna je funkcija enote in v kakšnih kontekstih se pravzaprav uporablja:

Takoj sem sklenil, da ga bom kupil, in prav zanima me, kdaj bo z njim hotela delati tudi moja ***boljša polovica****.*

Nobena ženska ni tako slaba, da bi ne mogla postati moževa ***boljša polovica****.*

Drugi, zame najpomembnejši argument, se pravi moja sredica, pa je, da ***boljša polovica*** *v vlogi popolne gospodinje enostavno izgubi spoštovanje v partnerski zvezi.*

Z Jonasom se je spoprijateljil v Mišolovki, kjer je igrala tudi njegova ***boljša polovica*** *Sabina.*

Kar izstopa, je dejstvo, da se iztočnica pojavlja skoraj izključno v imenovalniku, pogosto ob apoziciji (24 od 200 konkordanc), tj. občnem ali lastnem imenu. Nanaša se lahko na oba spola, vendar močno prevladuje raba, kjer zveza označuje ženske partnerice, najpogosteje žene, v humornem tonu. V 14 primerih gre za opis v križanki, torej formulaično, stereotipno rabo v specifičnem besedilnem žanru. V pregledanem vzorcu 200 konkordanc koligacijsko izstopajo svojilne modifikacije, najpogosteje zaimka *moja* (20) in *njegova* (38), pogoste pa so tudi *moževa* (13), redkeje *njena* (2), *tvoja* (2), *vaša* (4), in po enkrat *nesojena, očetova*, pri čemer gre samo v dveh primerih za žensko. Tipična je modifikacija bodisi s svojilnim zaimkom, svojilno obliko pogosto lastnega imena (*Robertova, Spielbergova, Tončkova, ravnateljeva*) ali ustrezna rodilniška struktura (*Magnifica, para, dua, tandema, zakoncev*), ki določa pripadnost. Te frazeološke enote ni mogoče razumeti na podlagi sestavinske zgradbe in/ali skladenjskih razmerij med sestavinami. Če je v SSKJ ustrezno določena konotacija *šaljivo* – glede na sekundarne pomene te LE, ki si jih delijo slovenski govorci, kot posredno kažejo tudi korpusni zgledi – pa sobesedila po drugi strani kažejo še mnogo več: gre za kulturni evfemizem, ki odraža naravo družbenih odnosov med spoloma tako, da blaži obravnavo sicer občutljivih intimnih razmerij znotraj družbenega diskurza. Metadiskurzivnost posredno potrjuje tudi pogostost rabe v navednicah. Semantično prozodijo je mogoče opisati kot realizacijo določenih konvencionalnih predstav o razmerjih med spoloma v konkretni situaciji, tj. pozitivne lastnosti, tipično pripisane ženski – čustveni in moralni primat, ki pa je že isti hip konvencionalno degradiran skozi humor ali dojemanje izraza kot zastarelega.

2.2.3 Semantična prozodija – avtomatsko pridobivanje podatkov

Če ne gre za enačenje s semantično preferenco, je posebnost in težava avtomatskega prepoznavanja semantične prozodije predvsem v tem, da se kolokacijski in

koligacijski vzorci ne porajajo nujno iz ponavljajočih se oblik, temveč lahko le iz ponavljajočih se pomenskih odtenkov (Philip 2009: 14) ter v dejstvu, da gre za pomen, ki nastaja hkrati na vseh ravneh razširjene enote pomena. Poskusi avtomatskega označevanja prozodičnih vsebin so bili praviloma izvedeni ob predpostavki, da gre za vrsto konotativnega pomena oz. podkategorijo semantične preference. Kot že navedeno, gre v resnici za pomenotvorne mehanizme, ki niso ulovljivi v formalne kategorije in večinoma niso eksplicitno izraženi s konvencionalnimi izraznimi sredstvi. Ker smo večkrat poudarili, da je med formalno oprijemljivejšimi kategorijami ključnega pomena koligacija, ki sicer zajema široko polje (slovnično-pomenskih) jezikovnih pojavov, prav tako pa tudi prepoznavanje besedilnih zvrsti in zunajjezikovnih situacij, se v Tabeli 1 osredotočamo prav nanjo. Koligacije je namreč mogoče razmeroma uspešno tudi avtomatsko prepoznavati, znotraj orodja Corpus Architect/SketchEngine (Kilgarriff et al. 2004) najbolj sistematično s t. i. direktivo CONSTRUCTION+UNARY (Krek in Kilgarriff 2006). S tem ukazom je mogoče avtomatično generirati opozorila, kot npr. *pogosto zanikano*, *pogosto v 3. os. ednine*, torej tipične podatke o koligacijah. V Tabeli 1 navajamo koligacijske in besedilnozvrstne kategorije, ki smo jih evidentirali v postopku ročne korpusne analize gradiva in ki so se – v medsebojnem prepletu in v razmerju z drugimi elementi – izkazali kot pomembni kazalniki prozodične funkcije:

Tabela 1: Pregled semantičnoprozodičnih kazalnikov na osnovi ročne analize 250 gesel LBS in nabora dodatnih gesel.

Besedna vrsta/ Kategorialna lastnost	**koligacije**				**LE v LBS**
glagol					
oseba	1. oseba	2. oseba	3. oseba	brezos.	plezati (preko/skozi)
spol	m. spol	ž. spol			
prehodnost	nepreh.	preh.	trpnik		enačiti
povratnost	si	se			
čas					
tip dopolnil/ določil	prislovna d.	predložna d.			bobnati (pri/v)
naklon	povedni	velelni	vprašalni		
	trdilnost	nikalnost			enačiti
modalnost	želelnost	nujnost/ obveza			plezati preko/skozi, enačiti
polariziranost	pozitivno	negativno			
faznost itd.					bobnati (pri/v)

Besedna vrsta/ Kategorialna lastnost	**koligacije**				**LE v LBS**
pridevnik					
skl. funkcija	povedna raba	prilastkovna raba			situacija
primerjava, stopnja					kislo jabolko
tip dopolnil/ določil	predložna	okoliščinska			nagubano čelo
prislov					
primerjava, stopnja					v prihodnje/bodoče
samostalnik					
sklon	imenovalniški	neimenovalniški			boljša polovica
število	nav. v množini				nečednost
priredja	vezalno	ločilno	apozicija		boljša polovica, kislo zelje
količinski izrazi					kislo zelje
modifikacija	postmodifikacija	premodifikacija			kislo jabolko, boljša polovica
VSI					
premi govor					
navednice (metadiskurzivna in metaenonciativna raba)					čas kislih kumaric, kislo jabolko
ločila					kislo vreme
pozicija – končna					
pozicija – vmesna					
odvisnik	vsebinski	okoliščinski			
govorno dejanje					
členitev po aktualnosti	tema	rema			
vrste besedil oz. diskurza// žanri // zunajjezikovni dejavniki					
leposlovje					
stvarna literatura					situacija
strokovna besedila					situacija
besedilni žanr	križanka	kuharski recept itd.			boljša polovica, kislo zelje

3 SKLEP

V prispevku smo se v kontekstu razprave o novem slovarju slovenščine osredotočili na t. i. periferne pomene, ki so intuiciji tipično nedostopni, določata pa jih kontekst in raba. Takšne (pragmatično opredeljive) pomene je bolj smiselno, včasih tudi edino mogoče, opisovati v leksikografski razlagi (npr. zgoraj navedeni *enačiti (kaj s čim), boljša polovica* itd.). Kadar gre pri določenem pomenu za zaznamovano rabo, vendar je to rabo težko opredeliti v tradicionalnih okvirih jezikovne in besedilne zvrstnosti, predlagamo identifikacijo okoliščin oz. pragmatiko situacije, ki ustvarjajo to zaznamovanost, in njihovo vključitev v sam pomenski opis (po potrebi v kombinaciji z oznako). Na podlagi analize gesel, izdelanih v LBS, in nabora dodatnih LE ugotavljamo, da pri velikem delu zlasti frazeoloških enot v SSKJ, ki nosijo oznako *ekspr.*, najverjetneje ne gre za konotacijo, temveč semantično prozodijo, torej pragmatično funkcijo LE v sobesedilu oz. razširjeni enoti pomena. Lahko zaključimo, da je mogoče prvo, kot jo razumemo v ožjem smislu (inherentnost), relativno uspešno označevati s kvalifikatorji (gl. 2.1), kontekstualno pogojene pomene, zlasti semantično prozodijo, pa je smiselneje in z vidika sporazumevalnosti pogosto edino mogoče predstaviti v pomenskem opisu (gl. 2.2).

Sistematično označevanje semantične prozodije, ki ga je celostno omogočila šele korpusna analiza, tj. natančnejše navajanje okoliščin in odnosa govorca do vsebine povedanega, torej vrednotenja, bi moralo postati del ustaljene prakse v leksikografskem procesu in predmet prizadevanj pri nadgradnji avtomatske obdelave besedil tudi in še posebej pri načrtovanem novem slovarju slovenščine. Kljub temu, da so bile v tujem, zlasti angleškogovorečem okolju, pa tudi domačem prostoru že opravljene posamezne raziskave, ki izpostavljajo pomen semantične prozodije za leksikalno analizo, je bil ta vidik pomena celo v najnovejših leksikografskih projektih deležen premalo pozornosti.

Oznake: slovarska baza in slovar

Iztok Kosem

Abstract
This paper discusses the various possibilities concerning dictionary labels that have been brought about by recent developments in lexicography, especially with regards to the content of the planned dictionary of contemporary Slovenian. First, key decisions concerning dictionary labels are presented. This is followed by a discussion on different possible approaches to labelling, including the use of automatic methods to identify label candidates. The main part of the paper focuses on labels in the proposed dictionary of contemporary Slovenian, with some options considered and suggestions provided on how to improve and optimise the labelling process and the subsequent visualisation of labels in the dictionary.

Keywords: labels, Slovenian, automatic methods, dictionary, user

Ključne besede: oznake, slovenščina, avtomatske metode, slovar, uporabnik

1 UVOD

Oznake so različni tipi slovarskih pojasnil, ki uporabnike opozarjajo, da ima beseda, besedna zveza, njen pomen itd. določene slovnične omejitve, da se nanaša na določeno časovno obdobje, tip besedila, regionalne posebnosti, da se uporablja na določenem strokovnem področju, da izraža določen odnos do vsebine ali udeležencev ipd.

V sodobni leksikografiji smo s korpusi dobili možnost zelo podrobne analize realne jezikovne rabe, z novimi slovarskimi mediji pa možnost prikaza večje količine informacij, ki jih lahko uporabniku predstavimo na različne načine. Vse to pomeni, da lahko informacije, povezane z oznakami, določamo sistematično za več elementov slovarske mikrostrukture, npr. za posamezne kolokatorje ali nize kolokatorjev, pa tudi v različnih oblikah, npr. v obliki opozoril, daljših pojasnil ipd. Pomembna sprememba v sodobni leksikografiji se je zgodila tudi v razmerju slovarska baza – slovar, saj sodobne slovarske baze vsebujejo veliko več informacij kot iz njih izpeljani slovarji, poleg tega pa lahko ena sama slovarska baza služi za izdelavo različnih slovarjev. Slovarska baza tako vsebuje informacije, relevantne za različne tipe slovarskih uporabnikov. To je pomembno tudi z vidika načina določanja oznak in njihovega prikaza v slovarju, saj je treba pri načrtovanju novih slovarjev – zlasti slovarjev, ki se jezikovnega opisa lotevajo povsem na novo, kot je predlagani slovar sodobnega slovenskega jezika (SSSJ) – ob upoštevanju potreb vseh možnih uporabnikov natančno opredeliti razmerje med slovarsko bazo in slovarjem.

V prispevku najprej predstavimo ključne odločitve glede uporabe oznak v slovarju, nato različne načine določanja oznak, vključno z možnostjo vključevanja avtomatskih metod, ki jih omogočajo sodobne jezikovne tehnologije. V osrednjem delu se osredotočimo na oznake v SSSJ z vidika razmerja med slovarsko bazo in slovarjem ter predstavljamo nekaj možnosti in predlogov, kako izboljšati oz. optimizirati procesa določanja oznak in njihovo vizualizacijo v spletni različici slovarja.

2 OZNAKE

V zvezi z oznakami moramo pri snovanju slovarja sprejeti več odločitev. Najprej se moramo odločiti, kaj želimo v slovarju označevati. Največkrat se posebej označujejo slovnične, stilistične, področne, časovne, regionalne in registrske posebnosti, nekateri slovarji, zlasti slovarji za tuje govorce, označujejo tudi pragmatične posebnosti in frekvenco. Od naštetih vrst označevanja so samo frekvenčne oznake

in v določeni meri tudi časovne tiste, za katere se lahko naknadno, torej med izdelavo slovarja ali celo po njej, odločimo, da jih bomo dodali.

Po izbiri vrste označevanja moramo določiti oznake za vsako izbrano vrsto. Z leksikografskega vidika je bolj učinkovito, če imajo leksikografi na voljo omejen nabor oznak za posamezno vrsto, saj morajo pri analizi gradiva v programu za izdelavo slovarjev samo izbrati ustrezno oznako iz nabora in se jim ni treba ukvarjati s poimenovanjem ali razmišljanjem, ali uporabiti obstoječo oznako ali oblikovati novo. Vendar je pri marsikateri oznaki nabor zelo težko vnaprej povsem omejiti (gl. razdelek 2), sploh pri izdelavi povsem novega slovarja. Pri določenih vrstah oznak je težava tudi stopenjskost oz. razlike med stopnjami, ki jih opredeljujejo posamezne oznake, npr. pri registrskih *formalno, manj formalno, zelo formalno, neformalno* ipd. Zelo pomembno je tudi vprašanje poimenovanja posamezne oznake, ki mora biti uporabniku jasno in razumljivo. Za SSKJ je bila recimo zaradi prevelike splošnosti in premajhne obvestilnosti kritizirana oznaka *ekspresivno*[1] (Müller 2009). Poleg tega so raziskave (Rozman 2010) med šolskimi uporabniki Slovarja slovenskega knjižnega jezika (SSKJ) razkrile tudi napačno tolmačenje okrajšanih oznak, vendar so tovrstne težave dejansko bolj ali manj omejene na tiskane izdaje slovarjev, kjer so zaradi prostorskih omejitev pogosteje uporabljane okrajšave.[2]

Ne glede na medij pa mora biti v slovarju jasno opredeljen način prikazovanja oznak z vidika njihove umestitve v slovarsko geslo. Po eni strani se odločamo, kam oznako pozicionirati, pred element, ki ga označujemo, ali za njim. To je na ravni slovarske mikrostrukture različno: medtem ko oznake sledijo iztočnicam, stalnim zvezam in frazeološkim enotam, jih pri pomenih in podpomenih ponavadi najdemo pred razlagami, tj. na samem začetku. Mesto oznake v geslu pa določa tudi njen doseg oz. opredeljuje, katere dele gesla oznaka zajema. Za leksikografa to ne predstavlja bistvenih težav: oznaka za iztočnico (v zaglavju) velja za celotno geslo, oznaka pred pomenom za celoten pomen itd. Vendar pa, kot opozarjata Atkins in Rundell (2008: 231), se takšne dosledne rabe oznak slovarski uporabniki niti ne zavedajo. Ker uporabniki enojezičnih slovarjev največkrat ne preberejo celotnega gesla, ampak se osredotočijo le na določene dele, največkrat na razlago, zapis iztočnice, sinonime in zglede (gl. Harvey & Yuill 1997; Hartmann 1999; Kosem 2010; Verlinde in Binon 2010; Lorentzen in Theilgaard 2012), se lahko, če je oznaka v zaglavju in velja za celotno geslo, pri obsežnejših geslih zgodi, da jo uporabnik prezre. Idealne rešitve ni, vendar pa digitalni mediji vsekakor ponujajo več možnosti različnih vizualizacijskih rešitev (o tem več v razdelku 3).

1 S tega vidika je presenetljivo, da so se avtorji Osnutka koncepta za novi slovar slovenskega knjižnega jezika (Gliha Komac et al. 2015) odločili obdržati to oznako.

2 Okrajšave najdemo tudi v elektronskih različicah slovarjev, a gre praviloma za prvotno tiskane slovarje, prenesene na splet ali v kateri drugi digitalni medij, kar velja tudi za SSKJ, SSKJ2 in SNB.

Od medija neodvisna vizualizacijska težava je kopičenje oznak. V takšnem primeru je treba uporabnikom jasno sporočiti, ali gre za razmerje in-in oz. ali-ali. Poleg tega je daljši niz okrajšanih oznak za uporabnika še bolj zahteven za tolmačenje, saj mora poznati pomen vsake okrajšave v nizu:[3]

> **bêdro** -a stil. -ésa s (é, ẹ́) noga nad kolenom; *stegno*: smejal se je in se tolkel po bedrih / obirati kurje bedro
> // nav. mn., pog., šalj. *noga*: hlače mu kar opletajo po suhih bedresih

Dodatno zahtevnost vnašajo kombinacije okrajšanih oznak in ostalih elementov, recimo številke za osebo pri glagolih, ki jih uporabnik ne najde na seznamu razvezanih oznak:

> **bíti**[1] bíjem nedov., 3. mn. stil. bijó; bìl (í î)

Če so že takšni primeri rabe okrajšanih oznak potencialno zahtevni za uporabnika, gotovo podobno velja za gesla, v katerih se oznake pojavljajo pri več (pod)pomenih in dejansko imajo pomembno vlogo pri razlikovanju med posameznimi (pod)pomeni ali pri prepoznavanju ustreznega (pod)pomena:

> **beséda**[2] -e ž, rod. mn. stil. besedí (ẹ́)
> 2. ed. in mn. **misel, izražena z besedami:** /…/
> 3. nav. ed., ekspr. zagotovilo, obljuba: /…/
> ed., star. dogovor: /…/
> 4. **izražanje misli z govorjenjem:** /…/
> 5. ed. in mn. **govorni ali pisni nastop v javnosti:** /…/
> // nav. ed. **možnost, pravica do govorjenja, zlasti v javnosti:** /…/
> 6. ed., knjiž. **izmenjava mnenj, misli;** pogovor, govor: /…/
> 7. ed., nav. vznes. **sistem izraznih sredstev za govorno in pisno sporazumevanje;** jezik: /…/

Za snovanje uporabniku prijaznejših rešitev imamo več možnosti, od uporabe različnih oblik pisav, različnih barv ali barvnih odtenkov itd. Glede uporabe barv npr. raziskave kažejo, da barvne oznake uporabniki hitreje opazijo, pa tudi informacijo si bolje zapomnijo (Dziemianko 2015). Razlikovanje lahko vzpostavimo tudi z dodelitvijo stalnega prostora (na zaslonu) za določene tipe oznak v geslu (gl. razdelek 3) in podobnimi vizualizacijskimi rešitvami.

Pri oznakah se moramo odločiti tudi o tem, kdaj sploh uporabiti oznako. Gre za vprašanje, kateri način podajanja informacije, ki jo ponuja oznaka, je z vidika uporabnika najustreznejši pri danem pomenu, podpomenu ali katerem drugem delu

3 Vsi primeri na tej strani so iz spletne različice SSKJ2.

gesla. Gantar in Kosem (2013) navajata dve obliki označevanja: eksplicitno in implicitno. Eksplicitno označevanje je klasično označevanje z oznakami, pri katerem je oznaka jasno ločen (največkrat enobesedni) element slovarske mikrostrukture. Pri implicitnem označevanju pa informacija, ki bi jo sicer posredovala oznaka, postane del slovarske razlage, bodisi zaradi manjšega izpostavljanja, tesnejše povezanosti z razlago (zlasti pragmatika) ali zaradi mejnosti (terminološko – splošno). Implicitno označevanje je dejansko lahko učinkovitejše od eksplicitnega, zlasti za materne govorce, saj raziskave kažejo, da uporabniki skoraj vedno preberejo razlago, zelo redko pa ločeno podane slovnične informacije in informacije o rabi (Hartmann 1999; Kosem 2010). V razlagi oznaka postane sestavni del pojasnjevalne informacije o pomenu besede, kar si uporabniki lažje zapomnijo (Barnbrook 2002), medtem ko je pri eksplicitnem načinu podajanja ločena od ostalih delov gesla in jo morajo uporabniki tolmačiti skupaj z razlago, zgledom, stalno zvezo ali s katerim drugim elementom slovarskega gesla. Pri implicitnem označevanju ima leksikograf tudi možnost uporabe daljše ubeseditve, kar je pri informacijah o pragmatičnih in stilističnih posebnostih rabe besede zelo uporabno.

Obstaja tudi kombinacija eksplicitnega in implicitnega označevanja, pri katerem oznako oz. informacijo, ki jo posredujemo, predstavimo v obliki komentarja za zgledi oz. proti koncu pomenskega opisa oz. razlage. McCreary (2004) je pri poskusih s študenti leksikografije ugotovil, da je takšna oblika označevanja z leksikografskega vidika lažja kot pa vključevanje informacije v razlago. Rezultat je informacijsko razbremenjena razlaga, hkrati pa leksikograf dobi možnost, da v komentarju napiše nekoliko izčrpnejše pojasnilo.

Pri označevanju gre torej predvsem za ugotovitev rabe, ki zahteva oznako, izbiro ustreznega načina predstavitve tovrstne informacije in pri eksplicitnem označevanju še za ustrezno vizualizacijo oznake v slovarju. V samem procesu morajo leksikografi odgovoriti na vprašanje, ali sploh uporabiti oznako, katero oznako uporabiti, in zlasti pri implicitnem označevanju, kako jo ustrezno ubesediti. V nekaterih primerih se oznaka sprva zapiše eksplicitno, v končni verziji slovarja pa je prestavljena v razlago. Kaj se v takšnih primerih zgodi s prvotno oznako? Jo sploh še lahko najdemo? Kako lahko v slovarju poiščemo razlage s takšnimi informacijami? Pri omogočanju takšnih funkcionalnosti igra ključno vlogo načrtovanje tega, kako se bodo oznake in z njimi povezane informacije določale na ravni slovarske baze in slovarja oz. slovarjev.

3 METODE DOLOČANJA OZNAK

Preden se posvetimo oznakam v slovarski bazi in slovarju, je treba nameniti nekaj besed metodologiji določanja oznak med analizo korpusnega gradiva. Sodobna

leksikografska analiza ni več zgolj ročna; razvoj jezikovnih tehnologij je namreč leksikografom omogočil tudi uporabo avtomatskih načinov pridobivanja podatkov o jezikovni rabi. Številni tipi oznak, npr. slovnične, registrske, področne in regijske, so namreč povezani s porazdelitvijo besed v korpusnih besedilih in jih je načeloma mogoče avtomatično pridobiti iz korpusa (Rundell in Kilgarriff 2011). Seveda to ne pomeni, da so oznake avtomatično pripisane v slovarsko bazo, ampak gre zgolj za opozorila na potencialne oznake, ki v bazo oz. slovar preidejo le, če jih potrdijo leksikografi.

Avtomatično pridobivanje slovničnih oznak je bilo že preizkušeno tudi za slovenščino, in sicer v zaključni fazi izdelave Leksikalne baze za slovenščino (LBS), ko so bila pri testiranju postopka avtomatskega luščenja leksikalnih podatkov (ALLP; Kosem et al. 2012b; 2013a) pridobljena tudi opozorila o potencialnih slovničnih oznakah. Osnovo za tovrstne informacije so predstavljale gramatične relacije v slovnici besednih skic v orodju Sketch Engine, za katere je mogoče glede na poizvedbo (besedo v iztočnici) skladenjske odnose definirati glede na en sam element, ki izpostavlja en pojav v odnosu do vseh ostalih elementov v korpusu. Na ta način lahko pridobimo podatek o tem, ali nek pojav, kot npr. množinska oblika, tretjeosebna oblika ipd. statistično izstopa. Za vsako od potencialnih oznak je treba določiti statistično mejo, pri kateri se oznaka pripiše v avtomatsko izvoženo geslo.

Medtem ko je slovnične oznake mogoče na tak način avtomatično izluščiti iz kateregakoli korpusa, pa to ne velja za druge vrste oznak. Če namreč želimo pridobiti podatek o npr. področju ali registru rabe iztočnice oz. posameznega pomena, morajo biti besedila v korpusu že opremljena z ustrezno informacijo. Tako je določeno besedilo ali celo njegov del (npr. odstavek) lahko opredeljeno kot *športno, na internetnih forumih* itd. Do takšnega korpusa pridemo tako, da izdelamo podrobno taksonomijo, za vsako taksonomsko kategorijo izberemo učno množico dokumentov, katerim taksonomske kategorije pripišemo ročno, potem pa s strojnim učenjem označimo vsa besedila (ali njihove dele) v korpusu.

V zvezi z določanjem oznak se moramo odločiti tudi o tem, ali bomo leksikografom ponudili izdelan nabor oznak za posamezen tip ali pa bodo leksikografi imeli že vnaprej relativno svobodo pri ubeseditvi oznak. Slednji pristop je primeren zlasti za tiste tipe oznak, »kjer jezikovna raba niha in kaže različne pomenske, stilne, pragmatične in druge omejitve, ki jih je težko ustrezno zajeti z vnaprej določenimi kategorijami« (Gantar in Kosem 2013: 145). Kot primer lahko navedemo bazo DANTE (Atkins et al. 2010; Rundell in Atkins 2011), kjer so takšen pristop uporabili pri oblikovanju pragmatičnih oznak. Končni nabor je vseboval 511 pragmatičnih oznak, a se jih samo 92 pojavi več kot enkrat. Pregled oznak pokaže, da se oznake pogosto malenkostno razlikujejo

v ubeseditvi, posredujejo pa isto informacijo (npr. tako *emphasis, emphatic* in *emphatic use* opozarjajo, da gre za poudarek; podobno *expresses disapproval* in *disapproval* opozarjata, da gre za neodobravanje), kar je pri izdelavi slovarja na podlagi baze dobro poenotiti. Po drugi strani fleksibilnost pri ubeseditvi oznak ponuja določene prednosti, kot je npr. razkrivanje stopenjskosti oznak, npr. *expresses disapproval* ('izraža neodobravanje'), *can express disapproval* ('lahko izraža neodobravanje'), *often expressing disapproval* (»pogosto izraža neodobravanje«), *expresses strong disapproval* (»izraža močno neodobravanje«). Pri vnaprej določenem naboru oznak bi namreč težko predvideli vse takšne odtenke v rabi.

Nekoliko drugačna možnost določanja oznak je uvedba njihove hierarhične strukturiranosti, kot jo npr. za področne oznake priporočata Atkins in Rundell (2008), za slovenski prostor pa Kosem (2011). Bistvo takšnega pristopa je, da hierarhija oznak z nadrejenimi in podrejenimi oznakami leksikografu ponuja možnost izbire splošnejše oznake (ali oznak), kadar se težko opredeli za eno samo (bolj določno). Če se leksikograf odloči za bolj določno oznako (npr. *tenis*), pa to hkrati pomeni avtomatičen pripis tudi njej nadrejene oznake oz. več oznak (npr. *šport*). Takšen način lahko kombiniramo tudi z opisanim fleksibilnim pristopom, kjer lahko splošnejše oznake določimo vnaprej, ubeseditev določnejših pa prepustimo leksikografom. To v fazi finalizacije slovarja omogoča hitrejše poenotenje oznak.

4 OZNAKE V SLOVARSKI BAZI IN SLOVARJU SODOBNEGA SLOVENSKEGA JEZIKA

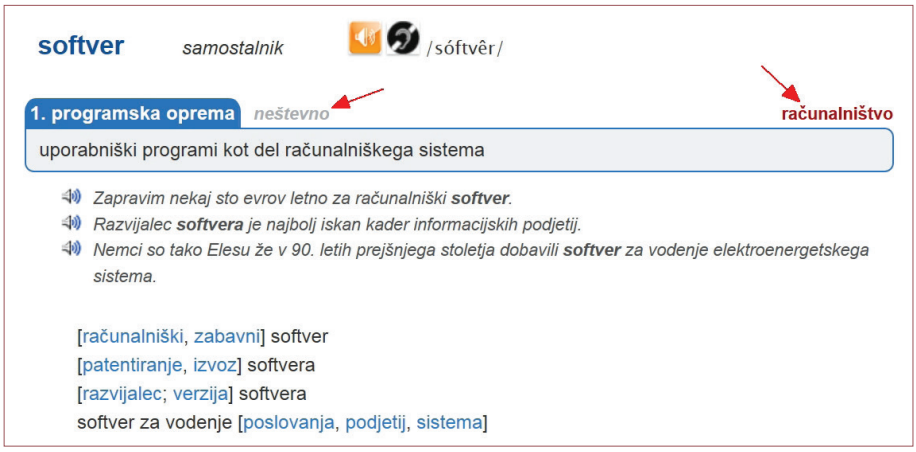
Prevladujoči medij sodobnih slovarjev je postal splet, ki ponuja možnost prikaza veliko večje količine informacij, tako tekstovnih kot multimedijskih, in omogoča različne povezave ter številne iskalne možnosti. Spletni slovarji »temeljijo na elektronskih slovarskih podatkovnih bazah, ki so strukturirane tako, da je podatke mogoče v čim večji meri pridobivati avtomatsko, jih urejati in povezovati z drugimi podatkovnimi bazami in uporabljati za nadaljnje jezikoslovne analize, hkrati pa jih izrabljati tudi v jezikovnotehnološke namene« (Gantar in Kosem 2013: 145). Ravno zaradi relevantnosti za raziskovalce in jezikovne tehnologe vsebujejo slovarske baze veliko več podatkov, kot jih vsebujejo na njej temelječi slovarji. Pri tem ne gre le za podatke, ki so primarno namenjeni računalniški obdelavi jezika, pač pa tudi za podatke, ki so koristni leksikografom pri izdelavi slovarskih gesel. O več slovarjih in ne o enem samem govorimo zato, ker je slovarska baza, zlasti slovarja večjega obsega, kot je SSSJ, lahko osnova številnim večjim in manjšim slovarjem, splošnim in specializiranim, za katere se predvideva, da bodo podatke v zvezi z jezikovno rabo vključevali različno.

Pri leksikografski analizi in izdelavi prvih različic gesel beležimo čim več z oznakami povezanih informacij. Sem sodi tudi avtomatsko luščenje potencialnih oznak, ki je opravljeno že pri izvozu podatkov iz korpusa. Pri beleženju oznak uporabljamo eksplicitno (oznaka) in implicitno metodo (razlaga oz. drugi del gesla), uporabljamo tudi t. i. skrite oznake, ki so eksplicitne oznake, a samo na ravni slovarske baze, saj jih uporabljamo pri omogočanju naprednejših iskanj po slovarju (glej razdelek 4.1). Leksikografi lahko skrite oznake uporabijo tudi v primerih, ko želijo opozoriti na informacijo, ki je sicer podana implicitno, in ko niso povsem prepričani o upravičenosti uporabe določene oznake.

Pri redakciji gesel se potrjujejo odločitve leksikografov, sprejete med analizo, in vnašajo morebitne spremembe, npr. eksplicitna oznaka postane skrita, informacija je posredovana implicitno. Na tej točki so tudi poenotene ubeseditve oznak. Med redakcijo je treba opravljati redne analize gesel z istimi oznakami ali z oznakami istega tipa, na podlagi ugotovitev pa se dopolnijo navodila za leksikografe in/ali izboljša postopek avtomatskega luščenja kandidatov za oznake, posledično pa se izboljša in pospeši analiza gradiva ter izdelava gesel.

Pri vizualizaciji oznak v slovarju se odločamo o oznakah, ki jih bomo na tak ali drugačen način vključili v slovar, in o tem, kakšno vlogo bomo vključenim oznakam namenili oz. kako, če sploh, jih bomo prikazali. Praviloma bodo v slovarju prikazane vse eksplicitne oznake v slovarski bazi, medtem ko bodo skrite oznake uporabljene večinoma zgolj za (naprednejša) iskanja (gl. 4.1). Če se slovarska baza uporabi za namene izdelave drugih slovarjev, pa se lahko zgodi, da se določene skrite oznake prikažejo tudi v slovarju ali pa se določena informacija v bazi preoblikuje v oznake, npr. informacija o frekvenci leme v oznake za skupine najpogostejših 1.000, 2.000 itd. besed v (pisnem) jeziku.

Čeprav je vizualizacija oznak zadnji korak v celotnem postopku, je eden najbolj ključnih. Že na začetku smo omenili nekaj pomembnih ugotovitev glede vizualizacije oznak, tako v zvezi z njihovo umestitvijo kot obliko (barvo ipd.). Elektronski medij nam ponuja številne možnosti, ki jih bo moral SSSJ čim bolj izkoristiti. Ena od možnosti prikaza oznak je predstavljena v Predlogu za izdelavo Slovarja sodobnega slovenskega jezika (Krek et al. 2013b: 34–35) in predvideva različne vizualizacijsek rešitve za različne tipe oznak (Slika 1). Prednost takšnega prikaza je, da ima vsak tip oznake točno določeno (prepoznavno) mesto in obliko (barvo, font itd.) v geslu oz. pri pomenu, ker predvidevamo, da bi to lahko uporabniku pomagalo informacijo hitreje opaziti in prepoznati ter jo zato tudi hitreje uzavestiti.



Slika 1: Prikaz slovničnih in področnih oznak (Krek et al. 2013b).

Pri snovanju vizualizacijskih rešitev je treba imeti v mislih različne pojavne oblike slovarja, od spletnega do slovarja na mobilnih napravah. Razlike v količini informacij, ki jih hkrati lahko prikažemo na različnih medijih, so namreč zelo različne in dejansko lahko govorimo o različnih tipih uporabnikov z vidika navad uporabnika medija.

V vsakem primeru morajo biti rešitve prikazovanja oznak in njihove ubeseditve preizkušene med slovarskimi uporabniki. Uporabniške študije se lahko začnejo izvajati že na začetku izdelave slovarja, saj se lahko uporabijo vzorčna gesla ali samo njihovi deli. Vpliv takšnih študij je zelo pomemben, saj rezultati vplivajo na vse korake v leksikografskem procesu, od določanja oznak do njihove vizualizacije.

4.1 Prednosti slovarske baze z vidika vključevanja oznak: nekaj možnosti

V tem razdelku predstavljamo nekaj možnosti, ki prikazujejo potencialno prednost uporabe ločenih korakov označevanja za slovarsko bazo in slovar za leksikografe in slovarske uporabnike. Naštete možnosti se bodo uporabile tudi pri izdelavi SSKJ.

4.2 Možnosti za leksikografe

Prednost slovarske baze kot podatkovnega vira, ki se razlikuje od dokončnega slovarja, je za leksikografe predvsem v bogastvu in raznolikosti informacij, ki jih

lahko vsebuje. Gre predvsem za podatke, ki so za dodajanje oznak dragoceni, bi jih pa pri tradicionalnem načinu določanja oznak prezrli oziroma bi jih upoštevali samo pri sprejemanju odločitev, ali oznako uporabiti ali ne. Vzemimo za primer slovnične oznake o številu, kot so *množina*, *v množini*, *navadno v množini*, *navadno v ednini*, *navadno v množini ali dvojini* itd.[4] Kot smo že omenili v razdelku 3, lahko opozorila o morebitnih slovničnih oznakah iz korpusa izvažamo avtomatsko, pri čemer za vsako oznako določimo določeno statistično mejo, ob kateri se opozorilo izpiše.[5] Vendar pa ob izpisu opozorila izgubimo informacijo o točnem odstotku pojavitev iztočnice v množini, ki bi bila koristna npr. za redaktorje pri potrjevanju geselskih člankov. Takšne informacije bomo pri izdelavi SSSJ vključili v bazo na ravni iztočnice in tudi na ravni posameznih kolokatorjev oz. skupin kolokatorjev ali celo struktur. Informacijo o pogostosti posameznih oblik iztočnice v korpusu Gigafida že imamo zapisano v Sloleksu (Dobrovoljc et al. 2015), tako da nam je v slovarsko bazo niti ni treba zapisati, ampak samo navedemo sklic na geslo v Sloleksu. Podatke o pogostosti pojavljanja oblik iztočnice s posameznim kolokatorjem pa lahko izvozimo med postopkom avtomatičnega pridobivanja podatkov iz korpusa. Takšni statistični podatki v slovarski bazi bodo zelo koristni pri posodabljanju gesel ob morebitnih povečanjih korpusa, saj bomo s pomočjo primerjave z novejšimi podatki lahko ugotovili morebitne spremembe v rabi iztočnice, njenih kolokatorjev in struktur ter posledično pomenov in oznake posodabljali oz. spreminjali veliko bolj sistematično. Dodatna prednost zapisovanja statističnih informacij v slovarsko bazo je tudi v tem, da lahko odstotkovni prag zelo znižamo ali celo izvozimo podatke za vse kolokatorje in strukture, potem pa se pri pripravi slovarja odločamo o mejnih vrednostih za vključitev posamezne oznake. Takšna rešitev je veliko boljša tudi za potencialne uporabnike slovarske baze, kot so npr. jezikoslovci, jezikovni tehnologi in drugi raziskovalci.

Ločena koraka analize in izdelave osnutkov gesel ter njihovega redakcijskega pregleda dajeta tudi možnost, da se poleg razlik v določanju oz. oblikovanju oznak vpelje različen pristop k določanju z oznakami povezanih informacij. Predstavljajmo si scenarij, da leksikograf pri analizi oblikuje razlago za določen pomen, ki vsebuje implicitno oznako, potem pa se redaktor odloči, da je bolje, da je oznaka predstavljena eksplicitno, kar seveda zahteva tudi preoblikovanje razlage. Del leksikografovega truda, vloženega v razlago, je tako izničen. Možna rešitev bi bila odprava implicitnega predstavljanja oznak oz. z njimi povezanih informacij v koraku analize in uvedba ločenega koraka (ali povečanje obsega nalog pri redakciji) za odločitve o dokončnem prikazu oznak v geslih. Pri analizi bi torej leksikografi oznake navajali samo eksplicitno, pri čemer bi morala uredniška ekipa slovarja pri vseh oznakah, ki bi jih lahko predstavili tudi implicitno, predvideti nekoliko

4 Navedeni primeri oznak so vzeti iz Leksikalne baze za slovenščino.

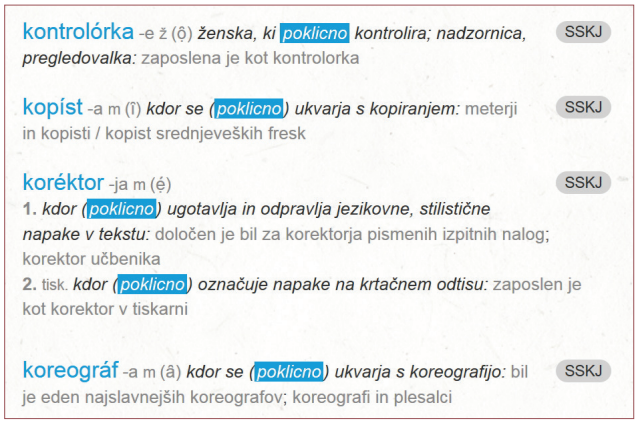
5 Mejne statistične vrednosti ne upoštevajo pogostosti določenega jezikovnega fenomena (npr. da je trpnik pri večini glagolov precej redkejši kot tvornik). Rundell in Kilgarriff (2011) svetujeta, da v takšnih primerih za kriterij vzamemo odstotkovno mero, npr. določen zgornji odstotek glagolov, ki se pogosto pojavljajo v trpniku.

večjo fleksibilnost pri ubeseditvi ali pa možnost uporabe krajše oblike oznake in daljšega pojasnila. Tako se tudi zmanjša možnost, da se implicitno predstavljena informacija o omejitvi rabe, pripisana med analizo, hkrati ne določi s skrito oznako in jo je kasneje v bazi težko najti.

V ločenem koraku ali med redakcijo bi se nato sprejemale odločitve o tem, kako najbolje predstaviti informacijo v oznaki. Takšna rešitev bi terjala tudi drugačen pristop k razlagam: med analizo bi se naredil neke vrste osnutek razlage za vsak (pod)pomen, lahko bi se uporabila tudi sinonimna razlaga ali kazalnik, med redakcijo pa bi se potem oblikovala dokončna razlaga.

4.3 Možnosti za slovarske uporabnike

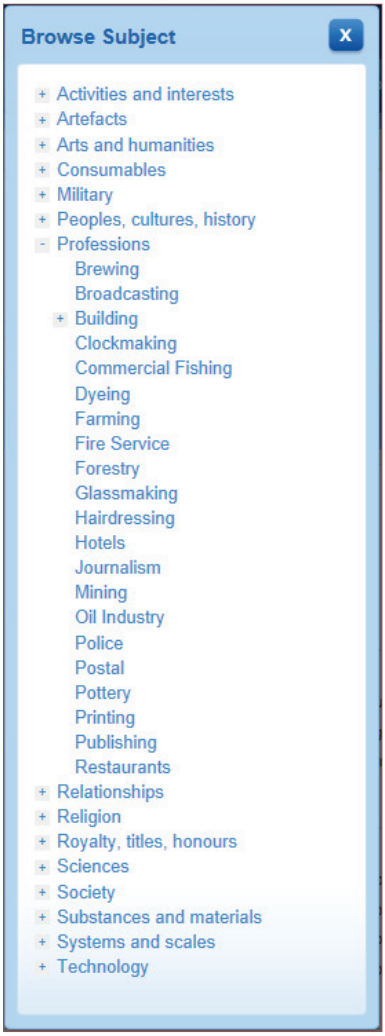
Pomembna prednost slovarske baze je predvsem v možnosti uporabe (določenih vrst) skritih oznak, ki so lahko zelo koristne za napredna iskanja. Pri tem lahko v bazo vključimo tudi tipe oznak, izdelane samo za slovarsko bazo. Prednosti takšne rešitve najlažje ponazorimo s primerom. Recimo, da želimo v slovarju poiskati vse iztočnice oz. pomene, ki pomenijo poklice. V obstoječih slovarjih, ki nimajo tako z informacijami bogatih slovarskih baz, moramo iskati določene besede v delih gesla, največkrat v razlagah. Tako lahko recimo pomene, ki se nanašajo na poklice, v SSKJ na portalu Fran poiščemo z iskanjem *poklicno* v razlagah (Slika 2).



Slika 2: Nekaj rezultatov za iskanje *poklicno* v razlagah SSKJ (portal Fran).

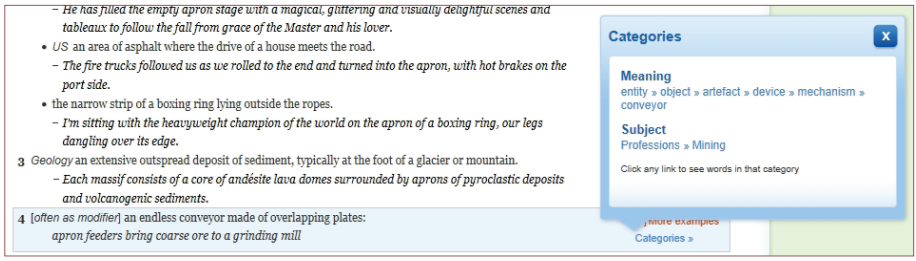
Težavi sta vsaj dve: po eni strani uporaba *poklicno* v razlagi ne pomeni nujno, da gre za poklic (npr. ***kri*** 3b *izraža socialno, poklicno pripadnost*), po drugi strani pa imamo iztočnice oz. pomene, pri katerih gre za poklic, a beseda *poklicno* v razlagi ni uporabljena (npr. ***prodajalec*** *kdor prodaja*).

V slovarski bazi to lahko rešimo tako, da pri vseh iztočnicah oz. njihovih pomenih in podpomenih, ki se nanašajo na poklice, uporabimo skrito oznako *poklic*, pa mogoče tudi z oznako določenega poklica. Tako lahko uporabnikom kasneje omogočimo, da na enostaven način poiščejo vse poklice ali vse iztočnice oz. pomene, povezane s posameznih poklicem. Podobno informacijo ima v slovarski bazi tudi angleški slovar Oxford[6] in jo je do nedavne prenove vmesnika ponujal pri naprednem iskanju (Slika 3), pa tudi v pojavnem oknu pri izbiri posameznega pomena (Slika 4).



Slika 3: Napredno iskanje po strokovnem področju v slovarju Oxford (prejšnja verzija vmesnika).

6 http://www.oxforddictionaries.com/ (dostop 8. 8. 2015).



Slika 4: Informacija o tematiki (Subject; v tem primeru o poklicu) pri določenem pomenu v slovarju Oxford (prejšnja verzija vmesnika).

Za naprednejše uporabnike, med katere štejemo npr. jezikoslovce in druge raziskovalce ter jezikovne tehnologe, bo dejansko bolj zanimiva slovarska baza kot sam slovar, saj njihove informacijske potrebe ponavadi presegajo informacije, ki jih ponuja slovar, poleg tega želijo čim večjo svobodo pri analizi in obdelavi jezikovnih podatkov. Za njihove potrebe lahko oznake v bazo dodamo tudi naknadno oz. neodvisno od leksikografske analize, npr. za označitev določenega besedišča, za katerega že imamo izdelan seznam. Na primer, za poučevanje in učenje slovenščine kot tujega jezika bi lahko uporabili oznake A1, B1, B2 itd. za označitev besedišča po ravneh skupnega evropskega referenčnega okvira (CEFR). Takšna informacija v slovarski bazi je potem koristna tudi za učitelje slovenščine kot drugega/tujega jezika, izdelovalce učnih gradiv, raziskovalce ipd., konec koncev pa tudi za leksikografe, ki bi se lotili izdelave slovarja za tujce.[7]

5 ZAKLJUČEK

V prispevku smo pokazali, katere korenite spremembe prinaša vpeljava slovarske baze, opremljene s številnimi informacijami, od katerih vse niso predvidene za prikazovanje v slovarju, tudi za postopke označevanja. Ob upoštevanju prednosti slovarske baze, kot so npr. fleksibilnost pri ubeseditvi oznak pri analizi korpusnega gradiva in možnost uporabe skritih oznak, nikakor ne smemo pozabiti na vlogo ustrezno pripravljenega korpusnega gradiva in raziskav med uporabniki, ki lahko še dodatno izboljša določanje oznak in njihovo ubeseditev ter vizualizacijo.

SSSJ bo izkoristil vse naštete prednosti za označevanje omejitev v rabi besed in podobnih informacij, saj bo le tako zadostil potrebam slovarskih uporabnikov, tudi naprednejših, navsezadnje pa tudi potrebam leksikografov, ki jim bo z uvedbo sodobnih postopkov določitve oznak in njihovega zapisa omogočil hitrejšo in objektivnejšo analizo korpusnega gradiva.

7 Glej prispevek Rozman et al. (2015) za diskusijo o tem, ali je potreben slovar za tujce ali pa bi njihove potrebe lahko zadovoljil tudi splošni slovar z določenimi dodanimi elementi.

IX
Slovar in slovnica

Besedne vrste v slovenskem jeziku

Robert Grošelj

Abstract
This paper presents a new part-of-speech system for the Slovenian language. The categorisation is based on three criteria: semantic, morphological and syntactic, which enable the identification of ten parts of speech: nouns, adjectives, verbs, adverbs, interjections, pronouns, numerals, prepositions, conjunctions and particles. In the first part of the study the part-of-speech criteria are described generally; the second part, however, is dedicated to a more specific (though still rather generalised) presentation of the above mentioned part-of-speech classes by taking semantic, morphological and syntactic criteria into account.

Keywords: Slovenian, morphology, lexicology, parts of speech, criteria

Ključne beside: slovenščina, morfologija, leksikologija, besedne vrste, kriteriji

1 UVOD

Problematika besednih vrst, ki se nahaja na meji med oblikoslovjem in leksikologijo (Adam 2015: 46), je gotovo področje, na katerega posega tudi koncept slovarja sodobnega slovenskega jezika (SSSJ) in, seveda, slovar sam, tako kot to velja za SSKJ in slovarski del SP. Ker omenjeni referenčni slovarski deli slonita na besednovrstnih členitvah iz 50. oz. najkasneje 70. let 20. stol. (prim. Bajec et al. 1956 za SSKJ; Toporišič 1974/1975 za SP)[1] in ker se je od takrat v slovenskem jezikoslovju zgodilo že marsikaj – med drugim so bile kritično ovrednotene same besednovrstne opredelitve, poleg tega pa je bilo opozorjeno tudi na njihovo neusklajenost v različnih slovarskih in slovničnih delih, tudi v okvirih iste besednovrstne teorije (prim. predvsem Toporišič 1988 za SSKJ; Krek 2015 za Toporišičevo slovnico, slovarski del SP in SSKJ2), se zdi smotrno besedne vrste v slovenskem jeziku ponovno celostno premisliti (prim. npr. razpravo Toporišič 1974/1975). Namen prispevka je tako podati novo členitev besednih vrst v slovenskem jeziku, ki bi lahko služila tudi kot izhodišče za slovarske aplikacije, obenem pa prispevati k sistematizaciji in uskladitvi te jezikoslovne problematike vsaj za del jezikovnih virov in opisov slovenskega jezika.

V razpravi bom skušal na podlagi treh kriterijev (pomenskega, oblikoslovnega in skladenjskega) opredeliti deset besednih vrst (samostalnik, pridevnik, glagol, prislov, medmet, zaimek, števnik, predlog, veznik, členek). Gre za členitev, ki se odmika od Toporišičeve besednovrstne teorije (z vodilnim skladenjskim kriterijem) in se približuje »tradicionalnemu« (v prvi vrsti pomensko-oblikoslovnemu) pogledu Bajca idr., hkrati pa upošteva spoznanja aktualnega slovenskega in neslovenskega jezikoslovja.

2 POTEK RAZPRAVE

Razprava v grobem sledi prikazoma besednih vrst v Cvrček et al. (2010: 128–133) in Štícha et al. (2013: 80–89). V prvem delu so predstavljeni in razčlenjeni trije besednovrstni kriteriji, in sicer tako, da so besedne vrste znotraj vsakega kriterija tudi umeščene oz. kontekstualizirane; na koncu poglavja so značilnosti besednih vrst in, posledično, razmerja med njimi predstavljeni še tabelarično. Izhodišče v drugem delu razprave so posamezne besedne vrste – njihova opredelitev vključuje vse tri kriterije (v zaporedju pomen, oblikoslovje, skladnja), v problematičnih primerih pa podajam tudi predloge za razdvoumljanje besednih vrst.

1 Podobno kot za SSKJ bi – ob manjših (kompromisnih) spremembah – veljalo tudi za SSKJ2 (prim. Krek 2015).

3 UVOD V BESEDNE VRSTE

Razlikovanje med besednimi vrstami ima v evropskem jezikoslovju več kot dvatisočletno, pri nas pa – z *Zimskimi uricami* Adama Bohoriča iz l. 1584 – več kot štiristoletno tradicijo (Linke et al. 2004: 79; Ahačič 2007: 118–166). Ne preseneča torej, da so se v času spreminjali tako kategorizacija besed, razumevanje izhodišč za njihovo členitev kot tudi njihovo število (za slovenščino npr. Černelič Kozlevčar 1988: 289–291; Toporišič 2003: 311–413; z vidika oblikoslovnega označevanja korpusov prim. Krek 2010: 69–109, 163–182), čeprav se – po mojem védenju – v večini sodobnih splošnih jezikoslovnih del še vedno pojavljajo prav različice »tradicionalnih« besednovrstnih opredelitev (prim. npr. Dardano in Trifone 1995; Morley 2000; Silić in Pranjković 2005; Cvrček et al. 2010). Domnevam, da ne brez razloga – gre namreč za ustaljeno kategorizacijo, ki se je skozi čas dopolnjevala in izpopolnjevala, se postopoma uveljavljala v opisih različnih jezikov, zaradi česar ima tudi močno kontrastivno-primerjalno vrednost (znanstveno, didaktično itn.).

4 BESEDNOVRSTNI KRITERIJI

»Tradicionalno« se besedne vrste delijo na osnovi treh kriterijev, tj. pomenskega, oblikoslovnega in skladenjskega, ki se med seboj prekrivajo in dopolnjujejo (npr. Bajec et al. 1971: 129; Dürscheid 2007: 20–23; Cvrček et al. 2010: 128–129; Salvi 2013: 16). Opredelitev besednih vrst v razpravi temelji na celostni umestitvi posamezne kategorije besed v okviru kriterijev, ki si sledijo v hierarhičnem zaporedju pomen, oblikoslovje in skladnja.

4.1 Pomenski kriterij

Najbolj abstraktno pomensko razlikovanje je med **polnopomenskimi** (avtosemantikami) in **nepolnopomenskimi besedami** (sinsemantikami). Razlikovanje v grobem izhaja iz tega, da imajo besede bodisi relativno samostojen (od drugih besed neodvisen) **leksikalni** pomen (predvsem samostalniki, pridevniki, glagoli in prislovi; sem so, kljub posebnim značilnostim, uvrščeni tudi medmeti) bodisi odvisen **slovnični** pomen (predlogi, vezniki, členki; tudi zaimki, števniki, a z drugačnimi značilnostmi; Čermák 2001: 179; Cvrček et al. 2010: 129; prim. Toporišič 1992: 190).[2]

Polnopomenske besede se delijo na poimenovanja substanc oz. entitet (samostalniki), njihovih značilnosti (pridevniki), dogodkov oz. dogajanj (glagoli) in na

2 Odločitev za termina **polnopomenske** in **nepolnopomenske** besede je načelne narave, čeprav se zavedam, da nekatere polnopomenske besede v določenih kontekstih izgubijo »polni pomen«, npr. *priti* v *priti v časopis*.

značilnosti dogajanj oz. značilnosti značilnosti (prislovi). Gre za nekakšen **sestavinski** pomen, saj besede označujejo sestavine zunajjezikovne realnosti (Čermák 2001: 179; Cvrček et al. 2010: 129; Adam 2015: 46).

Posebno vrsto predstavljajo medmeti, katerih pomen opredeljujem kot **globalno situacijski** – (čustveno, telesno) razpoloženje, namero ali zvoke (oglašanje) poimenujejo namreč tako, da situacije ne členijo na sestavine (npr. entitete, značilnosti, dogajanje; prim. Anward 2000: 728; Adam 2015: 47).

Slovnični pomen je lahko **razmerni** (relacijski) ali **nadomestni** (substitucijski). Z nadomestnimi besednimi vrstami, katerih pomen je določen sobesedilno (jezikovno ali zunajjezikovno), lahko substance/entitete, značilnosti nadomeščamo oz. nanje kažemo (zaimki), lahko pa izražamo tudi njihovo številskost (števniki). Razmerne besedne vrste, ki se pomensko-funkcijsko realizirajo šele v razmerju do drugih besed, izražajo medsebojna razmerja med sestavinami sporočanja (vezniki in predlogi – pri predlogih je vsaj na eni strani vedno substanca oz. entiteta) in/ali (sporočevalčevo) razmerje do neke sestavine sporočanja ali sporočila (členki; Čermák 2001: 179; Cvrček et al. 2010: 129; Adam 2015: 46–47).

Na podlagi pomenskega kriterija težko jasno razlikujemo med besednimi vrstami, med npr. predlogi in vezniki, vezniki in členki, zaradi česar je treba pomenski kriterij dopolniti z drugimi (prim. Dürscheid 2007: 26–27; Salvi 2013: 17–18).

4.2 Oblikoslovni kriterij

Po oblikoslovnem kriteriju se besede delijo na **pregibne**, glede na način oblikovnega spreminjanja oz. pregibanja (predvsem sklanjanje, spreganje, deloma stopnjevanje),[3] in **nepregibne**. Med pregibne besedne vrste sodijo besede, ki se oblikovno spreminjajo: *korak, koraka*; *lep, lepega*; *čakam, čakaš*; *lep, najlepši*.[4] S tovrstnim spreminjanjem ne dobimo novih besed, temveč le različne oblike osnovne besede (Toporišič 1992: 145; Cvrček et al. 2010: 129).[5] Nepregibne besede imajo eno obliko za vse funkcije (*dobro, pred, ker, tudi, fej*), lahko pa se spreminjajo zaradi glasovnih razlogov (prim. pri predlogu *s Katjo* in *z Marijo*; prim. Toporišič 2004: 413–414).

3 Stopnjevanje ni upoštevano kot odločilni oblikoslovni kriterij, saj zajema le del pridevnikov in prislovov; treba pa je poudariti, da lahko sodeluje pri določanju besednovrstne pripadnosti, prim. prislov *blizu – bližje – najbližje* (*komu*) in predlog *blizu* (*koga*). Prim. tudi sicer uvrstitev prislova med nepregibne besedne vrste (npr. Toporišič 1992: 226).

4 Vse pregibne besede se – površinsko – ne spreminjajo, nekatere namreč ostajajo v isti obliki oz. se pregibajo z ničtimi končnicami, npr. *mami, fejst* itn. (Toporišič 1992: 145; Cvrček et al. 2010: 129).

5 To pa npr. z leksikografskega vidika ali z vidika označevanja korpusov še ne pomeni, da ne morejo biti posamezne oblike – dogovorno – obravnavane kot samostojne iztočnice, pri čemer se na povezave z drugimi iztočnicami eksplicitno opozarja (prim. Dobrovoljc et al. 2015).

K pregibnim besednim vrstam se uvrščajo samostalniki, pridevniki, zaimki, števniki (se sklanjajo) in glagoli (se spregajo), ostale besedne vrste, tj. prislovi, predlogi, vezniki, členki in medmeti, pa so opredeljene kot nepregibne (Toporišič 2004: 256–257; prim. še Cvrček et al. 2010: 129).

Z oblikoslovnim kriterijem se lahko besedne vrste ločijo na nepregibne in pregibne, slednje pa na tiste, ki svojo obliko spreminjajo s sklanjanjem ali spreganjem. Takšen kriterij ne more razločevalno-kategorialno opredeliti nepregibnih prislovov, predlogov, veznikov, členkov in medmetov, med pregibnimi besednimi vrstami pa samostalnikov, pridevnikov, zaimkov in števnikov (prim. Cvrček et al. 2010: 129; Salvi 2013: 18–20).[6] Oblikoslovni kriterij, skupaj s skladenjskim, torej dopolnjuje pomenskega.

4.3 Skladenjski kriterij

Skladenjski kriterij temelji na vlogi in zvezah posameznih besednih vrst v stavku in besedilu (prim. Toporišič 1992: 273; Vidovič Muha 2000: 30–32; Dürscheid 2007: 24–25; Salvi 2013: 20–30). Polnopomenske besedne vrste lahko opredelimo glede na prednostne **stavčnočlenske** vloge (funkcije). Samostalniki so prednostno osebki in predmeti, pridevniki tvorijo s samostalniki besedne zveze, v katerih so ujemalni prilastki (lahko so tudi povedkova določila oz. povedkovi prilastki). Glagol je tipično povedek, medtem ko so prislovi prislovna določila (tudi prilastki pridevnikov). Medmeti, s svojim globalno situacijskim pomenom, tvorijo **samostojne stavčne ustreznike** (t. i. pastavke), redkeje imajo opredeljivo stavčnočlensko vlogo.

Nadomestni besedni vrsti, tj. zaimki in števniki, se stavčnočlensko obnašajo kot samostalniki in/ali pridevniki, tako da jih **skladenjsko težko razločevalno opredelimo**.

Pri razmernih besednih vrstah lahko – tako kot v primeru njihovega pomena – govorimo o **razmerni skladenjski** (in ne stavčnočlenski) vlogi. Predlog je v nadrednem razmerju do sledeče besedne zveze, ki predstavlja ali nadomešča substanco – določa ji sklon. Vezniki vzpostavljajo podredno ali prireDno razmerje med stavki in stavčnimi členi, pri čemer nimajo sklonskega vpliva, hkrati pa stojijo neposredno pred stavkom ali stavčnim členom, ki ga s prvim povezujejo. Tudi členki nimajo stavčnočlenske vloge, obenem pa jih **skladenjsko težko razločevalno opredelimo**;

6 Pri pregibnih besednih vrstah je treba upoštevati še slovnične kategorije in formalnooblikoslovne razlike, a z zadržkom – formalnooblikoslovne značilnosti (pripadnost posameznim sklanjatvenim vzorcem) in slovnične kategorije so lahko skupne različnim besednim vrstam (podvrstam), poleg tega pa izražanje slednjih sodi deloma v skladnjo (t. i. ujemalne kategorije; za češ. prim. opozorilo v Adam 2015: 48).

ti izrazniki različnih sobesedilnih in naklonskih razmerij so bodisi nestavčnočlenski del stavka (deloma položajsko premestljiv) bodisi samostojni stavčni ustrezniki (za celostno skladenjsko opredelitev slovenskih besednih vrst glej Toporišič 1974, 1974/1975; prim. še Jakop 2000/2001: 307–308; Adam 2015: 47–48).

4.4 Shematična predstavitev slovenskih besednih vrst

V spodnji predstavitvi slovenskih besednih vrst (kot vzor je služila tabela za češki jezik v Cvrček et al. 2010: 131) so vključeni kriteriji v že predstavljenem zaporedju pomen, oblikoslovje in skladnja, ki jih ponazarjajo posplošeni (poenostavljeni) opisi. Namen tabele je opozoriti na osrednje, bistvene značilnosti posameznih besednih vrst in tako ponuditi izhodišče za njihovo obravnavo v razdelku 5.

Tabela 1: Shematična predstavitev slovenskih besednih vrst

	kriteriji					
	pomenski		**oblikoslovni**			**skladenjski**
	polnopomenska	pomen	pregibanje	sklanjanje	stopnjevanje	vloga
samostalnik	da	substanca, entiteta	da	da	ne	osebek, predmet
pridevnik	da	značilnost substance	da	da	(da)	prilastek
glagol	da	dogajanje	da	ne	ne	povedek
prislov	da	značilnost dogajanja ali značilnosti	ne	ne	(da)	prislovno določilo
medmet	da	situacija	ne	ne	ne	samostojna
zaimek	ne/da	nadomestilo	da	da	ne	neenotna
števnik	ne/da	število	da	da	ne	neenotna
predlog	ne	razmerje s substanco	ne	ne	ne	zveza s samostalnikom,[7] vpliv na sklon
veznik	ne	razmerje med sestavinami sporočanja	ne	ne	ne	zveze stavkov in stavčnih členov
členek	ne	razmerje do sestavin sporočanja	ne	ne	ne	neenotna

5 PRIKAZ BESEDNIH VRST V SLOVENSKEM JEZIKU

Izhodišče v tem delu razprave je, kot že omenjeno, deset besednih vrst, tj. samostalnik, pridevnik, glagol, prislov, medmet, zaimek, števnik, predlog, veznik in členek. Njihova predstavitev upošteva vse tri besednovrstne kriterije, v težjih primerih pa je med drugim opozorjeno tudi na možnosti natančnejšega razločevanja med posameznimi vrstami.

5.1 Samostalnik

Samostalniki (substantivi, S) so polnopomenska besedna vrsta; poimenujejo predmete, osebe, živali, dejanja, stanja in značilnosti, in sicer kot neodvisne substance oz. entitete (npr. *miza*, *otrok*, *pes, ples, veselje, lepota*; prim. Cvrček et al. 2010: 130). Pomensko se členijo na posamezne razrede, krovno na lastna in občna imena (*Janez – potok*; Toporišič 1974/1975: 297; 2004: 275).[7]

Samostalniki so pregibni – lahko se namreč sklanjajo (čeprav skloni niso nujno površinsko oz. glasovno izraženi, prim. samostalnike tipa *a*, *mami*; Toporišič 2004: 257), pri čemer se v primeru razlikovalnih sklonskih oblik uvrščajo v različne sklanjatvene vzorce (oz. sklanjatve; pregledno Toporišič 1974: 30–33; natančno Toporišič 2004: 278–303).

Slovnične kategorije samostalnikov naj bi bile predvsem spol (s podspolom živo/neživo pri sam. m. spola), sklon in število (s številskostjo in številnostjo). V slovenščini so samostalniki še tretjeosebni in primarno določni (nedoločnost naj bi se izražala ločeno, npr. *en fant*; Toporišič 1974/1975: 297; 2004: 275–276).[8] Inherentna (notranja) sistemska kategorija je pri samostalniku zgolj spol (prim. Vidovič Muha 2000: 32), medtem ko se druge (npr. število, sklon, določnost) realizirajo šele v besedilu, tako da je njihova vrednost po Čermáku (2001: 126) zgolj besedilna oz. ustreza de Saussurjevi *parole*.[9]

Skladenjsko (stavčnočlensko) so samostalniki prednostno osebki in (brezpredložni) predmeti (Toporišič 1974/1975: 296); na teh skladenjskih položajih se namreč najbolje ohranjajo razlikovalne samostalniške značilnosti, pomenske in slovnične (celostna skladenjska opredelitev v Toporišič 2004: 273).[10]

7 Toporišič (1992: 93, 151) lastna imena člení nadalje na osebna (*Peter*), zemljepisna (*Slovenija*) in stvarna (*Litostroj*), medtem ko pri občnih navaja imena bitij (*človek*), stvari (*drevo*), snovi (*zemlja*) ali pojmov (*lepota*).

8 Po Vidovič Muha (1996: 121–122) gre za besedilno nedoločnost.

9 V nekaterih primerih je inherentna tudi kategorija števila (številskost), prim. enoštevilske samostalnike tipa *možgani*, *vrata*.

10 Prim. potencialni pomenski premik v smeri značilnosti substance in nerelevantnost kategorij spola, številnosti idr. pri samostalniku v funkciji povedkovega določila (Vidovič Muha 2000: 33); podobno bi veljalo za pomenski premik v smeri značilnosti dogajanja/značilnosti na položaju prislovnega določila.

5.2 Pridevnik

Pridevniki (adjektivi, Prid) so polnopomenska besedna vrsta in izražajo značilnosti substanc oz. entitet, tj. značilnosti oseb, živali, konkretne in abstraktne predmetnosti (npr. *lep, dolg, češnjev, sadni, bratov*; prim. Cvrček et al. 2010: 130). Pomensko jih lahko členimo na tri razrede: lastnostne oz. kakovostne (*lep, mlad*), svojilne (*očetov, Marijin*) in vrstne (*šolski, lanski*; Bajec et al. 1971: 157).[11]

Pridevniki so pregibni – se sklanjajo (čeprav skloni niso nujno površinsko oz. glasovno izraženi, prim. pridevnike tipa *roza, poceni*) in se v primeru razlikovalnih sklonskih oblik uvrščajo v manjše število sklanjatvenih vzorcev (Toporišič 1974/1975: 298; 2004: 321–325).

Del pridevnikov (t. i. lastnostni) se tudi stopnjuje (Toporišič 1992: 313; 2004: 325–326).

Slovnične kategorije pridevnikov so spol (s podspolom živo/neživo), sklon, število, stopnja in določnost (zadnji dve v delu pridevnikov), večinoma pa se realizirajo v besedilu. Kategorije spol, sklon, število so ujemalne kategorije, ki jih pridevniki načeloma izražajo v odvisnosti od samostalnika, medtem ko sta stopnja in določnost neujemalni (določnost se deloma kaže kot inherentna sistemska kategorija; prim. Toporišič 1974/1975: 298; 2004: 317).[12]

Pridevniki v stavku razvijajo predvsem samostalnike (tudi prek glagola), manj zaimke – skladenjsko (stavčnočlensko) torej nastopajo kot (levi) ujemalni prilastki v samostalniških zvezah (izrazito), povedkova določila in povedkovi prilastki (Toporišič 1974/1975: 298).

5.3 Glagol

Glagoli (verba, G) so polnopomenska besedna vrsta in poimenujejo dogodke oz. dogajanja, natančneje dinamično spremenljive značilnosti substanc (v času; prim. Čermák 2001: 182), npr. dejanje (*vrniti se*), stanje (*viseti*), potek (*bleščati se*) itn. (natančnejša pomenska členitev v Toporišič 2004: 345, 351–353).

11 T. i. količinski pridevniki (npr. Toporišič 2004: 318; prim. številske vrstne in količinske pridevniške besede v Vidovič Muha 1978: 269, 271–274) so uvrščeni deloma med številke deloma med prislove.

12 Po Toporišiču (2004: 317) določnost kot prosta kategorija zaznamuje le del (zaimenskih in nezaimenskih) pridevniških besed. Zaenkrat odprto ostaja vprašanje notranje določnosti npr. svojilnih in vrstnih pridevnikov. Kot inherentno določne jih pojmujeta Marušič in Žaucer (2007: 229; prim. še Toporišič 1992: 27–28), medtem ko Vidovič Muha (1978: 264, 269) meni, da ne izražajo kategorije določnosti (svojilni pridevniki, navaja avtorica, je ne izražajo s posebno končnico). Prim. tudi prispevek Vidovič Muha (1996: 122), po katerem govorne določnosti ne morejo izražati samo vrstni pridevniki, tvorjeni iz samostalnika, tipa *mestni* (*svet*), besedilno (primarno) določni so svojilni (*očetov, državen*), snovni pridevniki (*železen, čokoladen*) in pridevniki, tvorjeni iz krajevnih in časovnih prislovov (*zgornji, današnji*), nedoločni ali določni pa so lahko lastnostni pridevniki (*lep*, *globok*).

Glagoli so pregibni; od drugih pregibnih besednih vrst se ločijo po tem, da se spregajo – spreminjajo svojo obliko v osebi (Toporišič 1992: 300) oz. osebi, številu, času in naklonu (prim. Štícha et al. 2013: 450). Glede na razlikovalne sedanjiške ali nedoločniške oz. sedanjiško-nedoločniške pripone se razvrščajo v različne spregatvene vzorce, t. i. glagolske vrste (Toporišič 1992: 49; natančneje Toporišič 1974: 43–45; 2004: 360–388).

Glagolske slovnične kategorije so oseba, število, čas, naklon, način in vid (prim. Čermák 2001: 133–143; tudi Toporišič 1992: 48), deloma tudi spol (deležniki; prim. Toporišič 1992: 22; Cvrček et al. 2010: 236). Kategorija glagolskega vida se nahaja na meji med inherentno sistemsko kategorijo in leksikalnim pomenom (Cvrček et al. 2010: 245), medtem ko se ostale kategorije udejanjajo šele v besedilu (oseba, število in spol so ujemalne). Med glagolskimi kategorijami se sicer omenjajo še – bolj ali manj prekrivne – vezljivost (Čermák 2001: 134; Žele 2003a: 11), intenca oz. pomenska usmerjenost (Vidovič Muha 2000: 34) ali prehodnost (Toporišič 1974/1975: 303), ki jih lahko le pogojno obravnavamo kot slovnične, poleg tega pa niso izključno glagolske.[13]

Ob navedenem je treba poudariti, da se vse glagolske oblike ne spregajo (oz. ne izražajo kategorije osebe): spregajo se zgolj osebne glagolske oblike, neosebne pa ne.[14] Tako bi bilo mogoče pri glagolskih oblikah imeti za (prave) glagole le osebne glagolske oblike, deloma tudi deležnike, ki so vključeni v spreganje; pri neosebnih glagolskih oblikah se pojavlja možnost kompleksnejše kategorizacije na podvrste – tudi na podlagi konverznih prehodov. Pripadnost glagolu bi lahko kategorialno utemeljili na podlagi vida, sicer pa bi bilo možno – teoretično načelno – nedoločnike (*delati*, *jesti*) opredeliti kot posebne glagole-samostalnike (GS), namenilnike (*delat*, *jest*) in deležja (*sedeč*, *prepevaje*, *molče*) kot glagole-prislove (GP), deležnike (*delal, delala* ipd.) kot glagole-pridevnike (GPrid). Pripadnost drugim besednim vrstam (predvsem zaradi nabora drugih besednovrstnih značilnosti) izkazujejo glagolniki kot (iz)glagolski samostalniki in deležniki stanja na *-l, -n*/*-t* kot (iz)glagolski pridevniki (prim. Toporišič 1974: 47; 1974/1975: 303–304; 2004: 402–405; Černelič Kozlevčar 1988: 290, 295–296).

Tipična skladenjska (stavčnočlenska) vloga glagolov je povedkova (Toporišič 1974/1975: 302; prim. Cvrček et al. 2010: 130).

13 Po Žele je vezljivost pomenskoskladenjska (2003a: 11) oz. pomenska kategorija (2012: 105), podobno je po Vidovič Muha (2000: 34) »intenčnost glagolskega dejanja« kategorialna pomenska sestavina, medtem ko Toporišič (1992: 210) prehodnost opredeli zgolj kot »lastnost besed /…/, da se uresničujejo ob kakem predmetu /…/«. Čermák (2001: 134) po drugi strani meni, da je vezljivost lahko tudi posebna (čeprav ne le glagolska) oblikoslovna kategorija, če formalni signal vezljivosti (npr. predlog, verjetno tudi sklon) obravnavamo kot del besede v določenem pomenu.

14 Prim. Toporišičevo mnenje (1992: 48), da so glagoli v ožjem smislu le osebne oblike, v širšem smislu pa tudi deležniki, deležja, nedoločnik, namenilnik in glagolnik. Černelič Kozlevčar (1988: 295–296) navaja, da so pri glagolu »z besednovrstnega stališča sporne zlasti neosebne glagolske oblike, ki imajo oblikovne značilnosti drugih besednih vrst, vendar se po slovnični tradiciji zaradi pomenske povezanosti z glagolom obravnavajo pri njem«.

5.4 Prislov

Prislovi (adverbi, P) so polnopomenska besedna vrsta; predstavljajo predvsem značilnosti dejavnosti oz. dogodkov, ki jih izražajo glagoli (*hitro hoditi, priti včeraj*), ali značilnosti značilnosti, ki jih izražajo pridevniki, tudi prislovi (*zelo lep*; *izjemno natančno*; prim. Cvrček et al. 2010: 130). Pomensko se členijo na posamezne razrede, prim. okoliščinski prislovi – prostorski (*gori*, *vstran*, *povsod*), časovni (*včeraj*); svojstvenostni prislovi – lastnostni (*dobro*, *veliko*) in vzročnostni (*namerno*, *zaman*; Toporišič 2004: 407–409).[15]

Prislovi sodijo načeloma med nepregibne besedne vrste (Toporišič 2004: 406) – stopnjuje se namreč le del prislovov, in sicer lastnostni (prim. Toporišič 1992: 314), čeprav bi bilo treba sem vključiti še razmerne okoliščinske prislove tipa *blizu*, *pozno*.[16]

Skladenjsko oz. stavčnočlensko prislovi nastopajo – v prvi vrsti – kot prislovna določila (različnih pomenskih tipov), pa tudi kot prilastki pridevnikov in prislovov (Toporišič 1974/1975: 301; prim. še rabo v Toporišič 2004: 410–411).[17]

5.5 Medmet

Medmeti (interjekcije, M) so polnopomenska besedna vrsta – izražajo (čustveno, telesno) razpoloženje (*av, joj*), namero (*horuk*) ali posnemajo zvoke, npr. oglašanje (*mu, čof*; prim. Vidovič Muha 2000: 86–88; Toporišič 2004: 451),[18] pri čemer se njihov pomen lahko opredeli kot globalno situacijski, saj situacije ne členijo na posamezne sestavine.[19]

Medmeti so načeloma nepregibna besedna vrsta. Izjemo predstavljajo nekateri namerni, natančneje kontaktni medmeti, ki v omejeni meri »posnemajo« glagolsko spregatev (prim. Štícha et al. 2013: 86) – za vzpostavljanje stika s prejemnikom uporabljajo predvsem končnico 2. os. mn. (tudi dv.) velelnika (prim. *na, nata*; *lej, lejta, lejte*; prim. Karlík et al. 2002: 56).[20]

15 Prim. še drugačno pomensko členitev v Toporišič 1974/1975: 301. Pri Bajcu et al. (1971: 282–285) gre za krajevne, časovne, vzročne, načinovne in količinske prislove; prva dva razreda sta okoliščinska, naslednji svojstvenostni. – Vidovič Muha (2000: 36–37) prislove pomensko-skladenjsko deli na propozicijske (izražajo zunanjo okoliščino glagolskega dejanja – kraj, čas, delno vzrok) in modifikacijske (izražajo notranjo lastnost glagolskega dejanja – vrstni in lastnostni prislovi). Po Skubicu (1999: 229) imajo prislovi propozicijsko (okoliščinsko) funkcijo.

16 Toporišič (1992: 314) stopnjevanje prislovov uvršča celo v besedotvorje, prim. njegovo mnenje, da se lastnostni prislovi ne stopnjujejo, temveč so samo tvorjeni iz ustreznih pridevniških stopenj – npr. *dobro* iz *dober, bolje* iz *boljši, najbolje* iz *najboljši*.

17 Vidovič Muha (2000: 36) modifikacijske prislove označi tudi položajsko: lahko so samo obglagolski (*dobro*, *lepo pisati*) ali pa se razvrščajo ob vse besedne vrste, ki poimenujejo razsežnost oz. intenzivnost (*zelo jokati*, *zelo lep*).

18 Anward (2000: 728) medmete deli na ekspresivne (*av, joj*), direktivne (*pst, hej*), fatične (*mhm*) in deskriptivne (t. i. ideofone, npr. *čof, hrsk*).

19 Prim. celovito govorno dejanje kot denotat medmeta (Vidovič Muha 2000: 47).

20 Podobno uvrstitev ima SSKJ (prim. gesli *na*, *lej*), medtem ko Toporišič (1992: 101; 2004: 461) tovrstne prvine označi kot okrnjene glagole oz. glagolske okrnjence.

Medmeti so skladenjsko večinoma samostojne povedi, in sicer stavčni ustrezniki ali t. i. pastavki (Toporišič 1974/1975: 305), kar ustreza njihovemu globalnemu situacijskemu pomenu. Včasih imajo opredeljivo stavčnočlensko vlogo – nastopajo npr. kot povedki, pa tudi predmeti, prislovna določila (prim. Cazinkić 2012: 204).

5.6 Zaimek

Zaimki (pronomina, Z) so nadomestna (substitucijska) besedna vrsta; njihov pomen je določen šele sobesedilno (jezikovno ali zunajjezikovno) – z njimi lahko substance/entitete in značilnosti nadomeščamo oz. nanje kažemo. Pomensko jih lahko členimo na posamezne razrede: na osebne (*jaz, ti*; s povratno-osebnim zaimkom *se*),[21] svojilne (*moj, njen*; s povratno-svojilnim zaimkom *svoj*), kazalne (*ta, tisti*), vprašalne (*kdo, kaj*), oziralne (*kar, kateri*), nedoločne (*nekdo, nekateri*) in nikalne zaimke (*nihče*; prim. Bajec et al. 1971: 171; Cvrček et al. 2010: 210; drugačna in kompleksnejša razvrstitev v Toporišič 2004: 305–314, 335).

Zaimki so pregibna besedna vrsta – se namreč sklanjajo in se v primeru razlikovalnih sklonskih oblik uvrščajo v različne sklanjatvene vzorce (prim. Toporišič 1974: 33–35, 38; 2004: 305–314, 336–338). Glede na pomensko-nadomestno vlogo in skladenjske (tudi oblikoslovne) značilnosti jih lahko delimo na dve podvrsti: zaimke-samostalnike (ZS) in zaimke-pridevnike (ZPrid).[22]

Slovnične kategorije, ki jih zaimki lahko izražajo (gre za celostni nabor kategorij), so spol (s podspoloma živo/neživo, človeško/nečloveško), oseba, sklon in število (s številskostjo), pri čemer se posamezni razredi in podvrste zaimkov obnašajo kakovostno različno.[23]

Skladenjska (stavčnočlenska) vloga zaimkov je odvisna od besedne vrste, ki jo nadomeščajo; gre torej za skladenjske vloge samostalnikov ali pridevnikov, prim. 5.1 in 5.2.

21 Prim. tudi izvorno zaimenske *se/si, ga/jo* kot proste glagolske morfeme, ki so sestavni deli glagolov v vseh njihovih pomenih, kot v npr. *smejati se*, *odnesti jo*, *lomiti ga* (Žele 2001: 79; Gantar 2007: 89–91, 166; Krek et al. 2013: 73).

22 T. i. (iz)zaimenski prislovi (*nekje, vsakič*) so uvrščeni med prislove.

23 Vse tri spole ločijo ZPrid (ujemalna kategorija) in konverzni ZS, ostali ZS pa so načeloma m. in sr. spola ter ločijo podspola človeško in živo (*kdo – kaj*); izjema so osebni zaimki: 1./2. os. ed. in povratni osebni so brezspolski (*jaz, ti, sebe*), 1./2. os. dv./mn. so dvospolski (*mi*, *me*), 3. os. so trispolski (*on, ona, ono*). Zaimki so načeloma tretjeosebni, izjema so osebni zaimki za 1./2. os. in brezosebni povratno-osebni zaimek (svojilni in povratno-svojilni zaimki so tretjeosebni). Za sklon, prim. 5.1 (ZS) in 5.2 (ZPrid). Trištevilski so ZPrid, konverzni ZS (izjema je dvojinski *oba*) in osebni zaimki, povratno-osebni zaimek je brezštevilski, preostali pa so enoštevilski (edninski). Prim. tudi Toporišič (2004: 304).

5.7 Števnik

Števniki (numeralia, Š) so nadomestna (substitucijska) besedna vrsta – substance/entitete in značilnosti nadomeščajo tako, da natančno opredeljujejo njihovo količino (število). Pomensko jih lahko členimo na posamezne razrede, prim. glavne (*ena, pet*), vrstilne (*prvi, peti*), ločilne (*peter, petero*) in množilne števnike (*enojen*; prim. Bajec et al. 1971: 190; nekoliko drugačna členitev v Toporišič 2004: 329–330).

Števniki so pregibna besedna vrsta – se namreč sklanjajo; v primeru razlikovalnih sklonskih oblik se uvrščajo v različne sklanjatvene vzorce (prim. Toporišič 1974: 38; 2004: 330–331). Glede na pomensko-nadomestno vlogo in skladenjske (tudi oblikoslovne) značilnosti jih lahko delimo na tri podvrste: števnike-samostalnike (ŠS, npr. *milijon*), števnike-pridevnike (ŠPrid, npr. *dva, prvi, peter*) in števnike-pridevnike/samostalnike (ŠPrid/S, npr. *pet, petero*).[24]

Slovnične kategorije, ki jih števniki lahko izražajo (gre za celostni nabor kategorij), so spol, sklon, deloma število (s številskostjo) in določnost (oboje pri delu števnikov), pri čemer se posamezni razredi in podvrste števnikov obnašajo kakovostno različno.[25]

Skladenjske (stavčnočlenske) vloge števnikov so raznovrstne: glavna števnika *milijon, milijarda* (ŠS) imata skladenjske vloge samostalnikov (prim. 5.1), vrstilni, ločilni vrstni (petera vrata) in množilni števniki (ŠPrid) imajo vloge pridevnikov (prim. 5.2), enako pa načeloma velja še za glavne števnike 1–4 (prim. op. 27). Glavni števniki od 5 naprej in ločilni količinski števniki (*petero vrat*; ŠPrid/S) kažejo na kompleksnejše skladenjsko obnašanje.[26]

24 T. i. nedoločne števnike (npr. *več, toliko*) uvrščamo med (količinske) prislove, enako pa velja tudi za (iz)števniške prislove.

25 Vse tri spole ločijo ŠPrid (glavni števniki od 1 do 4 /tudi kot konverzni ŠS/, vrstilni, ločilni vrstni in množilni števniki), ŠS (*milijon, milijarda*) so enospolni, medtem ko imajo ŠPrid/S (glavni števniki od 5 naprej, ločilni količinski) poseben status – zaradi neujemalnosti (samostalniškosti) v im. in rod. (so pa spolsko enooblikovni). Vsi števniki se sklanjajo; pri ločilnih količinskih števnikih skloni niso površinsko oz. glasovno izraženi, pri višjih glavnih števnikih pa ne nujno – izjemi sta SŠ *milijon*, *milijarda* z glasovno izraženimi končnicami. Število naj bi poznali vrstilni, ločilni vrstni in množilni števniki (ŠPrid; prim. Toporišič 2004: 330), a ga poznajo tudi nekateri glavni števniki (trištevilska ŠS *milijon, milijarda* in vsaj dvoštevilski ŠPrid *en*, prim. *en človek, eni možgani*) – v slovenščini torej ne moremo posplošeno govoriti o števniškem številu (niti pri glavnem števniku) kot delu leksikalnega pomena (prim. Švedova et al. 1970: 334). Po Toporišiču (2004: 330; prim. Cazinkić 2012: 187) določnost kot prosta kategorija zaznamuje množilne in ločilne vrstne števnike, medtem ko Vidovič Muha (1978: 269, 271) navaja, da števniški vrstni pridevniki (vrstilni, ločilni vrstni in množilni) in določni količinski pridevniki (glavni, ločilni količinski in množilni števniki) ne izražajo kategorije določnosti. Na podlagi prispevka Vidovič Muha (1996: 122–123) bi lahko sklepali, da so števniki (količinski pridevniki) prvotno besedilno določni; prim. še poimenovanje *določni številnik* v Toporišič (1992: 28). Prim. Toporišič (2004: 329–330); Cazinkić (2012: 187).

26 Prim. raba (a) s štetim predmetom: v im. in tož. se obnašajo kot samostalniki – so števniško-samostalniško jedro besedne zveze (šteti predmet je v rod. mn.), v ostalih sklonih kot pridevniki – so števniško-pridevniško določilo (ujemalni prilastek) štetega predmeta v besedni zvezi; raba (b) brez štetega predmeta: v im. in tož. (slednji po analogiji s »pridevniško« rabo) zahtevajo ob sebi rod. mn. osebnega zaimka (prim. *Pet nas je prišlo. Videl jih je pet.*), v ostalih sklonih pa so neobvezno števniško-pridevniško določilo (ujemalni prilastek) os. zaimka (prim. *Darilo je dal petim. Darilo je dal vam petim.*). Slednje je značilno tudi za rabo (b) glavnih števnikov od 2 do 4 (števnik 1 puščam ob strani) v vseh sklonih. – Ko je v osebku besedna zveza s števniško-samostalniškim jedrom, je povedek v 3. os. ed. sr. spola. V zvezi s kategorijo osebe v prisojevalnih besednih zvezah tipa *Štirje ste pred menoj prebrali tale zapis – trije smo ga razumeli …* prim. razlago v Toporišič (2004: 608).

5.8 Predlog

Predlogi (prepozicije, Predl) so t. i. nepolnopomenska oz. slovničnopomenska besedna vrsta z razmernim pomenom, ki se funkcijsko realizira šele v razmerju do drugih besed – izraža namreč medsebojno razmerje med sestavinami sporočanja, navadno med glagolom in samostalnikom (*razmišljati o vremenu*), med samostalniki (*razmišljanje o vremenu*, *spomin na mladost*), samostojno s samostalnikom pa lahko tvori tudi jezikovne prvine za izražanje različnih okoliščin (*v nedeljo*; Karlík et al. 2002: 349).[27] Abstraktna pomenska razmerja, ki jih predlogi izražajo, lahko ločimo na dimenzionalna (kraj, čas) in nedimenzionalna (način, vzrok, posledica itn.; Silić in Pranjković 2005: 245).[28]

Predlogi so nepregibna besedna vrsta; pri oblikovni (izgovorni, pisni in izgovorni) variantnosti ne gre za pregibanje – varianta predloga je večinoma posledica sledečega glasovnega okolja, prim. *z* pred zvenečimi, *s* pred nezvenečimi glasovi (npr. *z letalom* – *s kolesom*; Toporišič 2004: 413–414).

Predlogi so stavčnočlensko nesamostojni oz. del stavčnih členov. Vežejo se s samostalniškimi zvezami (ali ustrezniki) v različnih sklonih (navadno brez imenovalnika; Toporišič 2004: 413) – izražajo podredno razmerje sledeče besedne zveze, funkcija podrednosti pa je vezana še na končnico oz. sklon (Vidovič Muha 2000: 29, 38). Predložne zveze (s predlogom kot jedrom) opravljajo različne stavčnočlenske vloge, npr. predmeta, prislovnega določila, prilastka.[29]

Slovenščina med drugim ne pozna t. i. osirotelih predlogov (Golden 2000: 5) – predlogi se navadno nahajajo neposredno pred svojim določilom, le redki lahko določilu sledijo (npr. *težavam navkljub*; prim. Silić in Pranjković 2005: 244).[30] Določilo predložne zveze je včasih lahko izpuščeno (*zgoraj brez*; Toporišič 1974/1975: 304).

5.9 Veznik

Vezniki (konjunkcije, V) so nepolnopomenska oz. slovničnopomenska besedna vrsta z razmernim pomenom, ki se funkcijsko realizira v razmerju do drugih

27 Prim. op. 7. Obstajajo še zveze s prislovom (*za kasneje*), glagolom (*oditi brez jesti in piti*, *imeti za povedati*) ali odvisnikom (neknjiž. *Odšel je, brez da bi pozdravil* – možnost večbesednega veznika *brez da*; Toporišič (1974/1975: 304)).

28 Po Čermáku (2001: 185) se pomensko običajno razlikuje med krajevnimi, časovnimi predlogi in predlogi, ki izražajo različna abstraktna razmerja. Za natančnejšo pomensko členitev predložnih zvez v slovenščini glej Toporišič (2004: 416–423).

29 V vezljivostni skladnji predlogi nastopajo kot (a) vezljivi predložni morfemi (pomenskoskladenjsko jih predvideva nosilec vezljivosti, večinoma glagol), in sicer (a.1) leksikalizirani (del nosilca kot leksema – predložnosklonska določenost, npr. *biti ob* 'izgubiti') ali (a.2) neleksikalizirani (del vezljivosti leksema – različna predložnosklonska določenost, deloma izbirnost, npr. *govoriti o* ali *bivati v*, *pod*, *sredi*), ter (b) družljivi predložni morfemi, ki jih nosilec ne predvideva, so torej del dopolnila (npr. *govoriti v sobi*; prirejeno po Žele 2003b: 153–159).

30 *Navkljub* je v SSKJ in SP opredeljen kot predlog samo v predodločilnem položaju (*navkljub težavam*), čeprav bi bilo treba zaradi sopomenskosti *težavam navkljub* = *navkljub (kljub) težavam* upoštevati oba. – Distribucijsko (položajsko) se lahko govori o prepozicijah in postpozicijah, ki skupaj tvorijo razred adpozicij (prim. Morley 2000: 43; Čermák 2001: 185).

besed – izraža medsebojno razmerje med sestavinami sporočanja, navadno povezuje dva samostalnika (*oče in mati*), pridevnika (*visok ter suh*), prislova (*lepo ali grdo*) ali stavka (*Ko smo vstopili, je ni bilo več*).[31] Abstraktna pomenska razmerja, ki jih med sestavinami sporočanja izražajo vezniki (prim. Cvrček et al. 2010: 131), se ločijo na posamezne tipe (prim. povzemalno Toporišič 1974: 49–50; Čermák 2001: 185; natančneje Toporišič 2004: 432–444).

Vezniki so nepregibna besedna vrsta in ne nastopajo kot stavčni členi (Toporišič 2004: 426; Cvrček et al. 2010: 131), čeprav so lahko njihovi deli (prim. osebek v *Oče in mati sta se vrnila*). Temeljna skladenjska (tudi besedilna) vloga veznikov je povezovalna (glej zgoraj), od predlogov pa se ločijo po tem, da ne vplivajo na sklon besednih zvez, ki jih povezujejo (prim. Toporišič 2004: 426). Obenem so vezniki tudi izrazniki temeljnih sintagmatskih razmerij – prirednosti in podrednosti (Vidovič Muha 2000: 38); priredna razmerja se vzpostavljajo med besednimi zvezami in stavki (*oče in mati*; *Sedimo in pišemo*), medtem ko podredna obstajajo načeloma le med stavki (*Rekel je, da pride*). Posebnost je podredni *kot* z medbesednozvezno vlogo (*Janez je starejši kot Marija*; Toporišič 1992: 351; 2004: 426).[32] Vezniki torej sotvorijo priredno oz. podredno zložene besedne (npr. samostalniške, pridevniške, prislovne) ali stavčne zveze, ki jih J. Toporišič (1992: 352) opredeljuje kot vezniške.[33]

Vezniki se pojavljajo neposredno pred skladenjsko sestavino, ki jo povezujejo (največkrat neprvo).[34] Na začetku zložene stavčne zveze se sicer v slovenščini (podobno kot v češčini) pojavljajo podredni vezniki (prim. Čermák 2001: 185; Toporišič 2004: 647).[35]

Pomensko (funkcijsko) se veznikom – v okvirih Toporišičeve vezniške besede – pridružujejo še sovezniki (nekateri zaimki in prislovi) in nekateri členki (Toporišič 2004: 426; prim. Čermák 2001: 185), ki jih je včasih težje besednovrstno razmejiti od veznikov (prim. predloge za razmejitev v Gorjanc 1998: 369–372). O veznikih glej tudi razpravo Pisanski Peterlin 2015.

31 Prim. veznike kot povezovalce besednih zvez in stavkov (Toporišič 1992: 351; Čermák 2001: 185), čeprav je lahko njihova vloga tudi širše besedilna (Gorjanc 1998).

32 Tip sintagmatskega razmerja je sicer tudi delitveno merilo za same veznike, prim. priredne in podredne veznike oz. konjunktorje in subjunktorje (npr. Silić in Pranjković 2005: 251).

33 V generativnem jezikoslovju vezniško zvezo zaznamuje vezniško jedro, dopolnilo in določilo veznika (prim. Golden 2000: 26, 37). S *stalno vezniško zvezo* je v razpravi sicer označen večbesedni veznik.

34 Pri veznikih je treba sicer upoštevati tudi ožje pojmovane stalne vezniške zveze (prim. frazeološki vezniki; Toporišič 1982: 367), ki so lahko enodelne (del ene od povezanih skladenjskih sestavin) ali dvodelne, ustrezneje večdelne (deli več sestavin). Naj na tem mestu opozorim le na kompleksnejše položajsko skladenjsko obnašanje večdelnih ali korelativnih zvez: (a) zveze pravih veznikov – pred vsako povezano skladenjsko sestavino se pojavlja veznik (npr. *Ali ne more ali noče*); (b) zveze veznika in soodnosnega izraza – obnašanje posameznih delov je manj enotno (prim. *če – potem, ne samo – ampak tudi*); uvrstitev med stalne vezniške zveze upravičuje vsaj en vezniški del (prim. Gorjanc 1998; Čermák 2001: 185; Toporišič 2004: 426–429).

35 Kar pa ne pomeni, da se nekorelativni (enodelni) priredni vezniki ne morejo pojaviti na začetku povedi – takrat signalizirajo medpovedna razmerja ali razmerja med delom besedila in njegovim nadaljevanjem (prim. Gorjanc 1998: 367; Čermák 2001: 185).

5.10 Členek

Členki (partikule, Č) so nepolnopomenska oz. slovničnopomenska besedna vrsta z razmernim pomenom, ki se funkcijsko realizira v razmerju do drugih besed (in sobesedilno) – izraža (sporočevalčevo) razmerje do neke sestavine sporočanja, tudi sporočila v celoti (gre torej za dodatno sporočanjsko-pragmatično sestavino), ali pa vzpostavlja pomenska razmerja v besedilu (Jakop 2000/2001: 310–315; prim. Čermák 2001: 185–186; Cvrček et al. 2010: 131).[36] V skladu s tem lahko členke funkcijsko razdelimo na dva osnovna razreda – na naklonske (*ali, žal, baje, predvsem*) in povezovalne (*namreč, torej*; Skubic 1999: 211; Žele 2014: 10–11).[37]

Členki so nepregibna besedna vrsta, mednje pa sodijo leksikalne enote različnega izvora in različnih funkcij (predvsem prislovi in vezniki; Toporišič 2004: 445; Žele 2014). Ker so zelo hominimni z drugimi besednimi vrstami in so besednovrstno težje »ulovljivi«, pogosto veljajo za »ostanek« med besednimi vrstami (prim. Čermák 2001: 185; Cvrček et al. 2010: 131).[38]

Členki niso samostojni stavčni členi, lahko pa nastopajo kot deli stavčnih členov (vsaj potencialno, npr. *Tam je bila tudi Jasna*) ali kot samostojni pastavki (npr. *Da, Ne*; Jakop 2000/2001; prim. Toporišič 1992: 17).[39]

Naklonski členki se po svoji modifikacijski funkciji približujejo prislovom (prim. Vidovič Muha 2000: 31, 36), skladenjsko pa se od teh ločijo ravno po svoji »nestavčnočlenskosti« (po njih se ne moremo vprašati; Toporišič 1992: 17). Distribucijsko se obnašajo različno – tako nekateri (potrjevalni členki, členki zanikanja, nesoglašanja, možnostni, verjetnostni členki in členki mnenja, domneve) nastopajo predvsem ob glagolih (kot del povedka, npr. *Vlak boš najbrž zamudil*), spet drugi (poudarni, izvzemalni, dodajalni in presojevalni) ob samostalnikih in izrazih količine (npr. *Ravno ti bi to moral vedeti*; Černelič 1991: 82; Jakop 2000/2001: 309).[40] Povezovalni členki se od funkcijsko sorodnih veznikov in prislovov (tipa *potem*, *najprej*) ločijo po položaju (nestabilen), odsotnosti kohezivne vezi in modifikaciji dela besedila, ob katerega se razvrščajo (Gorjanc 1998: 372; prim. tudi Skubic 1999: 217).[41]

36 Vprašanje je, koliko lahko členke v celoti opredelimo kot skrčke celih (izpustnih) stavkov (Toporišič 2003: 299–309; 2004: 445). Prim. še »modificiran izraz globinskega dela povedi« (Vidovič Muha 2000: 31). (

37 V prispevku Jakop (2000/2001: 310–315) se členki delijo na navezovalne, naklonske (naklonske v ožjem smislu in skladenjskonaklonske) ter poudarne (prva skupina predstavlja zgornje povezovalne, preostali skupini pa naklonske členke; op. R. G.). Za ostale členitve členkov v slovenskem jezikoslovju prim. Černelič (1991), Balažic Bulc (2015).

38 Pri členkih je treba omeniti še stalne členkovne zveze oz. frazeološke členke tipa *kje neki*, *kako da ne, tako ali tako* (Toporišič 1982: 367; Skubic 1999: 216–217).

39 Po samostojni pastavčni rabi se ločijo od npr. čeških členkov (prim. Cvrček et al. 2010: 131).

40 Prim. Toporišičevo ugotovitev (2003: 301), da se »samostojnost členkov nasproti prislovom /…/ kaže v njihovi premestljivosti glede na besedni red stavka«. Več o tem bi povedala položajska analiza.

41 Nekoliko problematični so skladenjskonaklonski členki v odvisnih vprašalnih, želelnih in velelnih stavkih (npr. *ali, naj*) in nekateri povezovalni členki s širšo sobesedilno (jezikovno in zunajjezikovno) razmerno vlogo (npr. *In kdo si ti?*). V zvezi s tem prim. poglede v Černelič (1991: 83); Gorjanc (1998); Skubic (1999: 217).

Prim. še glavno razliko med medmeti in členki: medmeti so izrazito skladenjsko samostojni (pastavki), med členki pa so izraziti stavčni ustrezniki redki (npr. *Da, Ne*); večinoma so del stavčne zgradbe (Černelič 1991: 83).[42] Za podrobnejšo obravnavo členkov glej še prispevek Balažic Bulc (2015).

6 SKLEP

V razpravi sem na podlagi pomenskega, oblikoslovnega in skladenjskega kriterija opredelil deset besednih vrst, tj. samostalnik, pridevnik, glagol, prislov, medmet, zaimek, števnik, predlog, veznik in členek. Tovrstna členitev se približuje – kot je bilo že omenjeno – bolj »tradicionalnim« besednovrstnim opredelitvam, ki imajo – vsaj po mojem mnenju – dobro aplikativno vrednost, tudi zaradi svoje umeščenosti v (zgodovinskih in sodobnih, sinhronih in diahronih) jezikoslovnih opisih različnih jezikov (tudi kontrastivno-primerjalno).

Pregledna besednovrstna členitev pa ne more biti popolna – jezik je namreč dinamičen fenomen (npr. Čermák 2001: 14), ki na besednovrstnem nivoju omogoča kompleksno prehajanje besed iz ene kategorije v drugo (npr. Toporišič 1992: 301–302). Čeprav razprava ni posegla na področje t. i. konverzije (sprevračanja), je treba ta pojav gotovo upoštevati pri natančnejši obravnavi besedja, kot se npr. pojavlja v slovarskih delih (v našem primeru SSSJ; prim. Krek et al. 2013b: 75–77). Kvalitativno obnašanje besede zaznamuje namreč nabor besednovrstnih značilnosti, ki se ob drugačnem obnašanju spremeni. In ravno upoštevanje tega bi predstavljalo jezikoslovno nadgradnjo (tudi aplikativno) prispevka – zaznati in ustrezno opredeliti vsakršno jezikovno obnašanje besede ob upoštevanju njegove statistične (kvantitativne) relevantnosti oz. stabilnosti. Pričujoča razprava pa je lahko osnovno orodje tudi za ta namen.

42 Žele (2014: 10) med besednima vrstama vzpostavlja v prvi vrsti pomensko razliko: členek spreminja pomen sporočila, medmet pa ga razpoloženjsko obarva.

Problematika veznika kot besedne vrste v enojezičnem slovarju

Agnes Pisanski Peterlin

Abstract

There is a strong tendency in lexicography to focus on lexical words at the expense of function words. Previous studies have identified specific issues that arise in dictionary definitions and descriptions of function words due to their specific characteristics, particularly their grammatical function and the restricted lexical meaning. The aim of this paper is threefold: to briefly present the conjunction as a part of speech, focusing on its definition, function and typology; to provide an overview of recent research on conjunctions; and to analyse the treatment of conjunctions in *Slovenski pravopis 2001* and *Slovar slovenskega knjižnega jezika*, highlighting potential challenges in the dictionary treatment of conjunctions in a monolingual general purpose Slovenian language dictionary.

Keywords: conjunction, function words, lexicography, monolingual dictionary

Ključne besede: veznik, funkcijske besede, leksikografija, enojezični slovar

1 UVOD

Glede na osnovno poslanstvo leksikografije ni presenetljivo, da je v lekikografskih raziskavah in v splošnih enojezičnih slovarjih standardnega[1] jezika poudarek predvsem na obravnavi **leksikalnih besednih vrst** (prim. npr. Osswald, v tisku), vprašanju obravnave **slovničnih oziroma funkcijskih besednih vrst** pa je namenjeno mnogo manj pozornosti. Adamska-Sałaciak (2008) opozarja, da se leksikografi tovrstnih besednih vrst pravzaprav bojijo (prim. tudi Kirkpatrick 1985: 11), Osswald (V tisku) pa poudarja, da je sama leksikografska obravnava slovničnih besednih vrst problematična. Sinclair (1991: 81) celo izrazi dvom, ali je slovnične besedne vrste res smiselno vključevati v slovar, ki je vendarle primarno namenjen obravnavi leksikalnih besednih vrst. Balažic Bulc (2009: 33) v tem kontekstu opozarja na »pomensko občutljive kategorije«, torej na tiste kategorije, pri katerih je pomen odvisen od konteksta, pri čemer je treba seveda poudariti, da je pomenska odvisnost od konteksta tudi sicer v ospredju sodobne leksikografije tudi pri obravnavi leksikalnih besednih vrst.

Splošni enojezični slovarji, namenjeni rojenim govorcem jezika, se s skladenjskimi lastnostmi funkcijskih besednih vrst ne ukvarjajo zelo podrobno, kajti predpostavlja se, da bodo uporabniki slovarja imeli vsaj osnovno znanje jezika, uporaba funkcijskih besed pa je del splošne gramatikalne kompetence (Osswald, v tisku). Analize različnih slovarjev pa pokažejo, da obstoječe parafraze pomena funkcijskih besed niso posebej uporabne (Hoekstra 2010: 1009; Coffey 2006: 159; Balažic Bulc 2009: 39) zaradi same narave posameznih funkcijskih besed.

Ker so slovnične besedne vrste tako heterogena kategorija, jih tudi specializirani slovarji funkcijskih besednih vrst obravnavajo zelo različno; vezniki gotovo sodijo med tiste slovnične besede, ki imajo najbolj kompleksne skladenjske lastnosti (Osswald, v tisku).

Namen pričujočega članka je na kratko predstaviti veznike kot besedno vrsto, in sicer z vidika njihove definicije, funkcije in tipologije, predstaviti novejše smeri jezikoslovnih raziskav veznikov, analizirati obravnavo veznikov v Slovenskem pravopisu 2001 (Toporišič et al. 2010), v nadaljevanju SP, in Slovarju slovenskega knjižnega jezika, v nadaljevanju SSKJ in SSKJ2, ter izpostaviti potencialne probleme obravnave veznikov v enojezičnem slovarju za slovenski jezik.

1 Izraz standardni uporabljam v skladu z izhodišči, ki jih izpostavlja Skubic (2003: 209–210), predvsem pa prispevek Gorjanc et al. (2015).

2 PROBLEMATIKA VEZNIKOV

2.1 Osnovne značilnosti veznikov

V slovničnem opisu so vezniki definirani kot nepregibna **slovnična besedna vrsta** (prim. Toporišič, 2011: 318; Quirk et al. 1992: 68, 72), ki nimajo vloge stavčnih členov (Toporišič 2011: 319), njihova funkcija pa je, da povezujejo stavke ali dele stavkov (Crystal 1995: 213); v slovenščini so nekateri vezniki (npr. *da*) izključno medstavčni (Toporišič 2004: 426).

V nasprotju s predlogi, ki imajo prav tako povezovalno funkcijo (Crystal 1995: 206), vezniki ne vplivajo na sklonsko obliko besed ali besednih zvez, ki jih povezujejo (Toporišič 2004: 426). Kot slovnična besedna vrsta nimajo vsebinskega pomena temveč slovnični pomen (prim. Vidovič Muha 2007: 400; Gramley in Pätzold 1992: 125), čeprav Gramley in Pätzold (1992: 125) poudarjata, da imajo vezniki vendarle pomembno leksikalno komponento (časovnost, vzročnost, dopustnost, pogojnost itd.).

Prvo temeljito obravnavo problematike veznika v slovenščini predstavlja Pogorelec (1964). Toporišič (2004: 426) uvršča veznike med **vezniške besede** (v njegovi tipologiji so to slovnične oziroma skladenjske besedne vrste); poleg veznikov med vezniške besede sodijo še oziralni in vprašalni zaimki ter nekateri členki. Toporišič (2007: 409) pa veznike opiše še kot **vezala** oziroma **junkcije**, med katere sodijo tudi **sovezniki**, to so prislovi, členki in zaimki, ki uvajajo odvisnike. Vezniki se po njegovem mnenju od soveznikov razlikujejo po tem, da so samo ali predvsem vezala, sovezniki pa so prvotno nekaj drugega »se pa rabijo tudi na začetku odvisnikov ali drugega dela priredja« (Toporišič 2007: 409) in imajo v nasprotju z vezniki funkcijo stavčnih členov. Ker vezniki in tako imenovani sovezniki opravljajo identično funkcijo, je zelo verjetno, da je pri slovarski obravnavi veznikov potencialno problematično prav razmejevanje med tema dvema kategorijama.

2.2 Vrste veznikov glede na obliko

Glede na obliko loči Toporišič (2004: 426) v slovenščini **enodelne** veznike (npr. *in, pa*) in **dvodelne** veznike (npr. *ne samo – ampak tudi*). Enodelni vezniki so po Toporišiču (2004: 427) lahko tudi **večbesedni**, vsi večbesedni vezniki razen *to je* so medstavčni. Večbesedni vezniki so lahko sestavljeni iz kombinacije prislova in veznika (npr. *zato ker*), iz členka in veznika (npr. *češ da*) ali iz predloga, kazalnega zaimka in *da, ko* ali *ker* (npr. *kljub temu da*). Dvodelni vezniki so lahko sestavljeni iz dveh enakih veznikov (npr. *ali – ali*) ali pa iz veznika in soodnosnice (npr. *tako – kakor*).

2.3 Vrste veznikov glede na razmerje med elementi, ki jih povezujejo

Glede na razmerje med elementi, ki jih veznik povezuje, veznike delimo na **priredne** (**parataktične**) in **podredne** (**hipotaktične**). Priredni vezniki povezujejo enote, ki imajo v povedi enakovreden status, npr. dva stavka, dve samostalniški besedni zvezi ali dva pridevnika (Crystal 1995: 213). Toporišič (2004: 432–433) za slovenščino navaja sedem vrst prirednih razmerij (vezalno, stopnjevalno, ločno, protivno, vzročno, pojasnjevalno in sklepalno).

Podredni vezniki povezujejo enote, ki v povedi niso enakovredne; tipičen primer je povezovanje glavnega in odvisnega stavka (Crystal 1995: 213). Po Toporišiču (2004: 433) je v slovenščini mogoče identificirati trinajst vrst podrednih razmerij (osebkovo, povedkovo, predmetno, krajevno, časovno, načinovno, primerjalno, posledično, pogojno, dopustno, vzročno, namerno in prilastkovno). Nekateri slovenski vezniki so večfunkcijski, kar pomeni, da se »uporabljajo v več vrstah priredja ali podredja« (Toporišič 2004: 434): primeri večfunkcijskih prirednih veznikov so *in, pa, ter* in *ali*, podrednih pa *da, ko* in *če* (Toporišič 2004: 435–436).

2.4 Novejše smeri v raziskovanju veznikov

V zadnjih desetletjih se v uporabnem jezikoslovju raziskave veznih elementov osredotočajo predvsem na povezave med deli besedila, torej **medpovedne povezovalce** oziroma **konektorje**. Raziskave konektorjev v veliki meri temeljijo na študijah kohezije v smislu diskurznih odnosov nad slovnično strukturo, kot sta kohezijo izvirno opredelila Halliday in Hasan (1976), kasneje pa je njuno delo sistematično nadgradil J. R. Martin (2003). Raziskovalci, ki se ukvarjajo s konektorji, navadno v prvi vrsti delijo konektorje na **medpovedne** in **medstavčne**: tako Halliday in Hasan (1976) razlikujeta med zunanjimi in notranjimi konjunkcijskimi odnosi, Van Dijk (1979), ki med konektorje šteje tudi medstavčne veznike, pa med semantičnimi in pragmatičnimi konektorji, pri čemer so slednji navadno na začetku povedi. Tudi v slovenskem jezikoslovju so se uveljavile različne definicije konektorjev: Žagar in Schlamberger Brezar (2009: 164), ki se s povezovalci ukvarjata v okviru teorije argumentacije, to kategorijo obravnavata kot del nadpovedne skladnje, podobno razmejuje med vezniki in vezniškimi konektorji Žele (2012), nasprotno pa Gorjanc (1998) med konektorje uvršča tako medpovedne kot medstavčne vezi; za natančnejši pregled razmejitve med vezniki in konektorji prim. tudi Schlamberger Brezar (2009). V besediloslovnih raziskavah se za slovenščino problematika veznikov in konektorjev pogosto prepleta s problematiko členkov (prim. npr. Smolej

2004), prav s problematiko obravnave konektorjev v enojezičnem referenčnem slovarju pa se ukvarja Balažic Bulc (2009).

Druge smeri raziskovanja veznikov segajo v žanrsko analizo (Rayson et al. 2001), na področje stilistike in foreznične lingvistike (Pavelec et al. 2008) in nenazadnje tudi kontrastivnega jezikoslovja v kombinaciji s prevodoslovjem. Prav v okviru slednjega so raziskave pokazale, da je tradicionalni slovnični pogled na veznike, ki izhaja iz enega samega jezika, včasih nekoliko preozek: identificirane so bile namreč velike razlike med jeziki (in posledično tudi v prevodih) celo pri rabi zelo pogostih veznikov, ki so na videz povsem neproblematični (npr. Ramm in Fabricius-Hansen 2005; Cosme 2006; Behrens 2008; Pisanski Peterlin 2010; Hirci in Mikolič 2014). Če upoštevamo, da so ena od pomembnih skupin uporabnikov slovarjev tudi prevajalci, bi morda prav izsledki kontrastivnih in prevodoslovnih raziskav na tem področju lahko služili za pripravo izhodišč za slovarski opis veznikov.

3 VEZNIKI V SP, SSKJ IN SSKJ2

Pregled obravnave veznikov v SP in SSKJ (ter SSKJ2) se osredotoča na vprašanje, katera gesla so v posameznem referenčnem viru opredeljena kot vezniki in kje med referenčnimi viri prihaja do razhajanj.

Analizirani so bili referenčni viri v elektronski obliki, ki je omogočala enostavno iskanje po ključnih izrazih (*veznik, vezniška, vez.*). Zabeleženi so bili vsi primeri rabe za posamezni vir. Sledila je primerjava naborov veznikov v izbranih referenčnih virih.

Primerjava naborov veznikov v SP, SSKJ in SSKJ2 pokaže nekaj neujemanja pri obravnavi veznikov. V tabeli 1 so predstavljeni tisti izrazi, ki so v SP obravnavani kot vezniki, v SSKJ in SSKJ2 pa ne.

Tabela 1: Gesla, ki so v SP obravnavana kot vezniki, v SSKJ pa ne.

	SSKJ	*SSKJ2*
navrh	prislov	prislov
ergo	prislov	členek
namreč	prislov v vezniški rabi	členek
odnosno	-	-
vezniške zveze *navzlic temu da, od kod, od koder, samo da, sedaj ko, še ko, šele ko, tačas ko, tako da, takoj ko, tako kakor, tako kot, tako da, takrat ko, toliko da*	-	-

Kot je razvidno iz Tabele 1, se v SP pojavljajo tri iztočnice, ki so opredeljene kot vezniki, v SSKJ in SSKJ2 pa kot prislov oziroma členek, to so *navrh* (v SP je tudi homonimna iztočnica *navrh*, ki je prislov), *ergo* in *namreč*, pri čemer je *namreč* v SSKJ opisan kot prislov v vezniški rabi. SP obravnava tudi veznik *odnosno*, ki se kot geslo ne pojavi ne v SSKJ, ne v SSKJ2. V SP se kot samostojna gesla pojavljajo nekatere vezniške zveze, ki v SSKJ in SSKJ2 niso obravnavane kot iztočnice (*navzlic temu da, od kod, od koder, samo da, sedaj ko, še ko, šele ko, tačas ko, tako da, takoj ko, tako kakor, tako kot, tako da, takrat ko, toliko da*), seveda pa so kot gesla obravnavani vsi elementi teh vezniških zvez; prav tako so vsaj nekatere od omenjenih vezniških zvez navedene v zgledih pri posameznih elementih vezniških zvez.

Na prvi pogled je neujemanje med SP in SSKJ/SSKJ2 večje, kajti SP kot veznike poleg tega obravnava vrsto izrazov (*drugače, magari, odklej, samo, tako, tedaj, vendar, vendarle* in *zato*), ki so v SSKJ in SSKJ2 opredeljeni kot prislovi oziroma členki, ena od njihovih rab pa je opisana kot »vezniška«, med tem ko so v SP ti izrazi homonimne iztočnice, od katerih je ena veznik, druga pa prislov, v nekaterih primerih pa je homonimnih iztočnic še več. Podobno SP kot veznik obravnava tudi *ki*, in sicer v smislu vzročnega veznika v primerih, kot so *To imaš, ki si ji dajal potuho* (tovrstna raba je opisana kot starinska); SSKJ in SSKJ2 sicer navajata isti primer in izpostavljata vzporednico z vzročnim veznikom *ker*, vendar takšne rabe eksplicitno ne opisujeta kot vezniške, jo pa seveda opredeljujeta kot starinsko. V teh primerih gre torej za razliko med obravnavo posameznega izraza z različnimi rabami (SSKJ in SSKJ2) in obravnavo izraza kot več homonimnih iztočnic (SP).

Zelo podobna situacija se pojavi pri dvodelnih veznikih (npr. *ali – ali*, *kakor – tako*): SP jih obravnava kot ločene iztočnice, SSKJ in SSKJ2 pa ne.

Podobne razlike je mogoče opazovati tudi pri primerjavi v drugo smer. V Tabeli 2 so predstavljeni tisti izrazi, ki so v SSKJ in SSKJ2 obravnavani kot vezniki, v SP pa ne.

Tabela 2: Gesla, ki so v SSKJ in SSKJ2 obravnavana kot vezniki, v SP pa ne-

primer	*SP*
nakar, odkar	prislov
kadar, kadarkoli, kakorkoli, kamor, kamorkoli, kar, kjer, kjerkoli, koder, koderkoli	prislovni zaimek
akopram, bilo, daravno, ino, jedva, ka, liki, nego, ni, no, pak	-
precej k, vtem ko	-

Kot je razvidno iz Tabele 2, se v SSKJ in SSKJ2 pojavita dva veznika (*nakar* in *odkar*), ki sta v SP obravnavna kot prislova. Zanimivo je, da SSKJ in SSKJ2 kot veznike obravnavata vrsto izrazov (*kadar, kadarkoli, kakorkoli, kamor, kamorkoli, kar, kjer, kjerkoli, koder, koderkoli*), ki so v SP opredeljeni kot prislovni zaimki. V SSKJ in SSKJ2 so ti izrazi večinoma homonimne iztočnice, od katerih je ena prislov in druga veznik, izjemi sta le *kar* (tri homonimne iztočnice, zaimek, prislov in veznik) in *kadar*, ki je ena sama iztočnica. V SSKJ in SSKJ2 se prav tako pojavi vrsta gesel, označenih s kvalifikatorji »zastarelo«, »starinsko« in »narečno« (*akopram, bilo, daravno, ino, jedva, ka, liki, nego, ni, no, pak*), ki jih v SP ni. V SSKJ in SSKJ2 se pojavita dve vezniški zvezi, ki v SP nista navedeni kot gesli, in sicer *precej ko* in *vtem ko*: v SP se sicer pojavi iztočnica *precej* (prislov) v pomenu takoj, ki je označena s kvalifikatorjem »zastarelo«, vezniška zveza *precej ko* pa ob njej ni posebej omenjena. V SSKJ in SSKJ2 so nekatere rabe drugih besednih vrst opisane kot »vezniške«, vendar tovrstne formulacije niso bile zajete v pričujočo analizo.

Razlike med SSKJ in SSKJ2 v obravnavi veznikov so sicer minimalne. Nekatere se navezujejo prav na že omenjeno »vezniško rabo«, ki je v SSKJ2 pri nekaterih izrazih opredeljena kot »členkovna raba« (raba prislova *dalje* v pomenu izražanja »obstoja nečesa poleg že povedanega« je tako v SSKJ opisana kot »vezniška«, v SSKJ2 pa »členkovna«, razlaga in zgledi pa so enaki). Podobno so posamezne rabe nekaterih veznikov v SSKJ2 opisane kot členkovne, v SSKJ pa niso posebej opredeljene, tako npr. *oziroma* v smislu »pravzaprav« *(šport oziroma alpinizem je gojil že od mladih nog*) ali so opisane kot ekspresivne, npr. *da* za izražanje ukaza ali želje (*da mi pri priči izgineš*), ali prislovne, npr. *saj* za izražanje ugotovitve (*torej je priznal. Saj mu ni preostalo nič drugega)*.

4 RAZPRAVA

Analiza iztočnic, ki so v SP in SSKJ ter SSKJ2 opredeljene kot vezniki, pokaže precejšnje konceptualno ujemanje pri obravnavi veznikov, kar je verjetno posledica dejstva, da gre za razmeroma jasno definirano besedno vrsto: vezniki povezujejo stavke ali dele stavkov (Crystal 1995: 213), v nasprotju s prislovi in zaimki nimajo vloge stavčnih členov (Toporišič 2011: 319) in v nasprotju s predlogi ne vplivajo na sklonsko obliko (Toporišič, 2004: 426).

Potencialno problematično je razmerje med vezniki in nekaterimi prislovi/členki ter redkeje zaimki, torej tistimi elementi, ki jih Toporišič (2007) imenuje »sovezniki« in uvajajo odvisnik ali drugi del priredja (Toporišič 2007: 409). Verjetno prav zaradi prepletanja funkcij v SP in SSKJ/SSKJ2 prihaja do nekaj neujemanja pri razlikovanju med vezniki in prislovi oziroma členki (*navrh, ergo, namreč, nakar* in *odkar*). Kriteriji za razdvoumljanje niso zelo natančno razloženi: »V bistvu

so vse to vezala (junkcije), vezniki pa so tista vezala, ko so samo ali predvsem to, sovezniki pa tiste besede, ki so prvotno kaj drugega, se pa rabijo tudi na začetku odvisnikov ali drugega dela priredja« (ibid.).

Poleg tega se pokažejo neujemanja pri obravnavi izrazov, ki imajo več slovničnih funkcij – SP in SSKJ/SSKJ2 se pri posameznih izrazih različno odločata glede tega, ali jih obravnavata kot več različnih gesel, ali kot eno geslo, znotraj katerega so opisani različni pomeni. Tako SP izraze *drugače, magari, odklej, samo, tako, tedaj, vendar, vendarle* in *zato* obravnava kot homonimne iztočnice, od katerih je ena samostojna veznik, druga pa prislov, SSKJ in SSKJ2 pa te izraze obravnavata kot prislove/členke, pri katerih je ena od rab vezniška (v Uvodu v SSKJ/SSKJ2, (2014: 28), je posebej izpostavljena težnja, da bi bilo homonimnih iztočnic čim manj). V Uvodu v SSKJ/SSKJ2 (Bajec in sod. 2014: 32), je navedeno tudi, da pri slovničnih besedah »odloča o razvrščanju razlag pogostnost«, pri čemer ni razloženo, kako je bila pogostnost izmerjena, »ob enaki pogostnosti pa ustaljeno slovnično zaporedje«.

Nasprotno pa se v SSKJ in SSKJ2 pojavljajo homonimne iztočnice *kadarkoli, kakorkoli, kamor, kamorkoli, kar, kjer, kjerkoli, koder, koderkoli*, od katerih je ena veznik in ena zaimek, ti izrazi so v SP opredeljeni kot prislovni zaimki. Podobna je tudi problematika *ki* (med veznike ga uvršča SP) in *kadar* (med veznike sodi po SSKJ in SSKJ2).

Na vprašanje ureditve slovarja in uvajanja samostojnih gesel se navezuje tudi obravnava dvodelnih veznikov, ki so v SP načeloma obravnavani kot ločena gesla, v SSKJ in SSKJ2 pa kot podgeslo.

Analiza pokaže tudi, da se zelo pomembno vprašanje odpira v zvezi z obravnavo vezniških zvez (v Toporišič 2004 so to »večbesednimi vezniki«): SP in SSKJ/SSKJ2 se glede tega, katere vezniške zveze je smiselno obravnavati kot ločena gesla, ne ujemata v celoti: korpusna analiza, narejena na osnovi skladenjsko razčlenjenega korpusa, bi utegnila pomembno prispevati k bolj sistematičnemu odločanju na tem področju.

Seveda se med referenčnimi viri pojavljajo tudi razlike pri vprašanju vključevanja posameznih veznikov (zastarelih, narečnih), vendar je odločitev o tem vezana predvsem na splošno usmeritev referenčnega vira.

Sklenemo lahko, da bi alternativna slovarska predstavitev morala upoštevati naslednja izhodišča:

- besednovrstna opredelitev funkcijskih besed je v nekaterih primerih problematična;

- v teh primerih se postavlja vprašanje, ali sodijo slovnične informacije v izhodišče opisa ali predstavljajo za uporabnika dodatne informacije;
- če je informacija o besednovrstni opredelitvi del slovarskega opisa, bi bilo smiselno zagotoviti, da je razmejitev med vezniki in prislovi, členki ter zaimki, ki uvajajo odvisnik oziroma drugi del priredja, sistematična in teoretično osnovana;
- kriterije za razvrščanje razlag pri homonimnih iztočnicah bi bilo treba opredeliti jasneje in natančneje kot v SSKJ/SSKJ2 (2014: 32);
- smiselna je specifična obravnava večbesednih enot (dvodelni vezniki, vezniške zveze, večbesedne enote, ki so v korpusu razpoznavne po ponavljajočih vzorcih in pri katerih je mogoče ugotavljati specifično besedilno funkcijo);
- slovarsko obravnavo bi kazalo prilagoditi tudi potencialno pomembni ciljni skupini slovarskih uporabnikov medjezikovnih posrednikih (npr. prevajalci, tolmači, učitelji tujega jezika).

5 SKLEP

Pričujoča analiza je pokazala, da je besednovrstna[2] opredelitev veznikov v slovarju razmeroma neproblematična, predvsem zaradi njihove jasne definicije. Analiza je prav tako pokazala, da ostajata odprti dve večji vprašanji: kakšna je optimalna slovarska obravnava tistih izrazov, ki združujejo funkcije prislovov/členkov (ali morda tudi zaimkov) in veznikov, ter katere vezniške zveze bi bilo smiselno obravnavati kot ločena gesla. S teoretičnega vidika bi razmejitev med vezniki in sorodnimi besednimi vrstami lahko temeljila na modelu, ki ga predstavlja Gorjanc (1998). Zdi pa se, da bi bila pri odločanju o teh vprašanjih lahko v veliko pomoč tudi korpusna analiza, ki bi dala podatke o pogostnosti rabe.

Analiza gradiva, ki je predstavljena v tem prispevku, se je osredotočala predvsem na vprašanje, kateri izrazi so bili v posameznih referenčnih virih opredeljeni kot vezniki. Vprašanje slovarske obravnave veznikov pa gotovo zajema tudi to, kakšni slovarski opisi veznika so za posamezne vrste enojezičnih referenčnih virov najbolj primerni. V Uvodu v SSKJ/SSKJ2 (2014: 31) je izpostavljeno dejstvo, da so slovnične besede obravnavane nekoliko specifično, saj imajo »v razlagah opisane svoje funkcije, sicer so obravnavane po enakih načelih kakor navadne besede«. Lang (1989) poudarja pomen vključevanja slovničnih spoznanj v geselski sestavek in meni, da je nujno, da se v opis veznikov vključijo sintaktični podatki in podatki o sintaktičnih omejitvah (povzeto po Osswald, v tisku). Vendar pa je je nadvse pomembno vprašanje kompleksnosti sintaktičnih

2 Mnogo bolj sta seveda problematični funkcijska in pomenska opredelitev.

podatkov, ki naj bodo v splošni enojezični slovar vključeni: kot opozarja Balažic Bulc (2009: 39), skladenjski opisi konektorjev v obstoječih jezikovnih virih za slovenščino zahtevajo »visoko jezikovno kompetentnost« in so za to namenjeni le strokovnjakom. Balažic Bulc (2009: 39) opozarja še na eno pomanjkljivost obstoječih skladenjskih opisov konektorjev: ne vsebujejo namreč informacij o tem, »kdaj in kako naj se določeno jezikovno strukturo uporabi v komunikaciji«. Razširitev raziskave na obravnavo veznikov v enojezičnih slovarjih standardnega jezika, namenjenih rojenim govorcem za različne jezike, in raziskava, ki bi se osredotočila na potrebe nekaterih najpogostejših skupin uporabnikov enojezičnega slovarja pri opisu veznikov, bi omogočili boljši vpogled v potencialne izboljšave slovarskih opisov.

Členek v slovenskem jezikoslovju in slovarju

Tatjana Balažic Bulc

Abstract
In Slovenian linguistics particle has been an independent word class since the 1970s, but there still is no general consensus on what particle actually is and which classification would be the most appropriate to encompass its wide range of meanings and functions. This is reflected in the fact that, as a word class, particle often serves as a »repository« for unclassified lexis. The aim of this paper is to present the previous research studies conducted on Slovenian particles. Different researchers have discussed particle from various theoretical perspectives: from syntactic-semantic approaches and a variety of functional-semantic theories, to text linguistics and pragmatics. However, this knowledge has for the most part not been implemented in the Slovenian lexicography. Despite various approaches to the issue of particles, some studies identify common characteristics that could serve as a starting point for further studies, which will certainly benefit from the findings in corpus linguistics.

Keywords: Slovenian, word classes, particle, particle classification, lexicography

Ključne besede: slovenščina, besedne vrste, členek, klasifikacija členka, leksikografija

1 UVOD

Čeprav je v slovenski besednovrstni kategorizaciji členek že od 70. let prejšnjega stoletja samostojna kategorija, v slovenskem jezikoslovju vse do danes ni nekega konsenza, kaj členek pravzaprav je. Njegovo opredelitev v precejšnji meri otežujeta pomenska oziroma, bolje rečeno, funkcijska neulovljivost jezikovnih elementov znotraj same besedne vrste in njegova pogosta homonimnost z drugimi besednimi vrstami, zlasti s prislovom in veznikom (več o tem npr. Krek 2015), dodatna oteževalna okoliščina pa so tudi nekatere druge lastnosti členka. Ena od teh je prav gotovo njegova »mejnost«, saj bi lahko rekli, da je členek oblikoskladenjska kategorija, ki se pomensko odraža na besedilni ravni. To dokazuje tudi dejstvo, da so členki v slovenskem jezikoslovju pogosto klasificirani kot jezikovni elementi, s katerimi tvorec besedila pravzaprav strukturira besedilo in sodijo na ravni besedila med metabesedilne elemente. Kot pravi Krek (2010: 222), je členek

> ena od težavnejših kategorij /…/. V osnovi gre za kategorijo, ki je besednovrstno gledano motivirana predvsem s pomenom in deloma s skladenjsko vlogo, ne z notranjo zgradbo pripadajočih elementov. Trenutno opredeljevanje sega od povsem skladenjske vloge stavčnih konektorjev do diskurznih elementov, ki imajo do neke mere pomensko določljivo vlogo (mja, mh, mda, naa itd.), kar kategorijo spreminja v slabo določljivo odprto množico elementov, pri katerih posledično postane nujno razlikovanje med odprto kategorijo členka, prislova (gotovo, ravno, resnično, približno, preprosto itd.), medmeta, samostalnika (hudiča, zlodja, jok, figa, drek itd.) in drugih kategorij.

V dosedanjih raziskavah je členek prikazan z različnih zornih kotov in tudi teoretična izhodišča njegovih klasifikacij so zelo raznolika. Podrobnejše preglede posameznih pristopov so predstavili že drugi avtorji (gl. npr. Černelič 1991; Jakop 2000/2001, Smolej 2001), v tem prispevku pa skušam strniti dosedanja preučevanja členkov in med različnimi pogledi na problematiko poiskati morebitne skupne točke, ki bi lahko nakazale smernice za nadaljnja preučevanja. V ta namen bo najprej podan pregled teoretičnih izhodišč pri posameznih avtorjih, temu pa sledi kratek prikaz, kako se teoretične predpostavke odražajo v slovenski leksikografski praksi, kjer se zahtevnost pomenskega opisovanja členkov pokaže v vsej svoji luči.

2 RAZLIČNI VIDIKI OBRAVNAVE ČLENKOV V SLOVENSKEM JEZIKOSLOVJU

Kot je že v uvodu omenjeno, se členek v slovenskem jezikoslovju kot samostojna besedna vrsta pojavi v 70. letih prejšnjega stoletja, ko ga 1974. leta v okviru

nove kategorizacije besednih vrst izpostavi Toporišič (1974/1975). Po njegovem mnenju tradicionalno razlikovanje devetih besednih vrst (samostalnik, pridevnik, zaimek, števnik, glagol, prislov, predlog, veznik in medmet) »takoj pokaže veliko šibkih točk«, zato predlaga:

> V slovenskem jeziku obstajajo 4 velike besedne (makro-) vrste: samostalniška in pridevniška beseda (obe sklonljivi), glagol (spregljiv) in prislov (nespregljiv, nesklonljiv, pač pa deloma pregiben); poleg tega je še 5 besednih vrst: členek, predlog, veznik, medmet in predikativ /…/ (Toporišič 1974/1975: 33).

Vendar pa je imel členek, kot ga danes razumemo, poseben status že pri njegovih predhodnikih, kjer v okviru prislova tvori dve samostojni skupini poudarnih in miselnih prislovov (gl. npr. Bajec et al. 1956). Kot pravi Černelič (1991: 82):

> Pregled obravnavanja členkov v slovenskem jezikoslovju kaže, da so primeri členkov registrirani od vsega začetka, vendar so obravnavani v okviru prislovov, ker nimajo posebnih oblikoslovnih lastnosti. Njihova posebnost med prislovi se je pokazala, zlasti ko so se začeli oblikovati kriteriji za določanje prislovov. Taka merila so: možnost, da se po prislovu vpraša; značilnost, da pojasnjuje glagol, pridevnik, prislov in le izjemoma samostalnik kot desni prilastek; in končno – v stavku so samostojni stavčni členi. Morfološke značilnosti, vprašalnica, skladenjska vloga in pomen pa so osnovna merila za razločevanje besednih vrst.

Podobnega mnenja je tudi Toporišič, ki v svoji slovnici (1976: 192, enako tudi 2000: 255) pravi: »Besedne vrste so v tej knjigi obravnavane kot pojmi za množice besed z enakimi skladenjskimi vlogami in drugimi lastnostmi (npr. tvorjenost, slovnične kategorije, konverzivnost ipd.). Skladenjske in druge lastnosti posamezne besedne vrste morajo biti razločevalne.« S skladenjskimi kriteriji Toporišič (1974/75: 39) utemelji tudi novo besedno vrsto:

> V sodobnem jezikoslovnem pojmovanju je členek besedna vrsta, ki se od prislova loči – preprosto povedano – po tem, da nima vprašalnice, tj. ne daje obvestila o kraju, času, načinu itd., ampak le o odnosu do vsebine povedi ali njenega dela, in torej tudi ni stavčni člen. Take besede so v naši tradicionalni slovnici imenovani poudarni in miselni prislovi /…/.

2.1 Tipološka klasifikacija členkov glede na skladenjsko-pomenski kriterij

Toporišič členke opredeli kot nepregibno besedno vrsto, ki se po vlogi včasih približa veznikom ali prislovom. Kot pravi, s členki »vzpostavljamo zveze s sobesedilom, izražamo pomenske odtenke posameznih besed, delov stavka, celih stavkov

in povedi ali pa tvorimo skladenjske naklone« (Toporišič 1976: 384). Njihova besednoredna pozicija je »pred tistim delom stavka, ki ga posebej poudarjajo« (ibid.: 540). Kasneje opredeli tudi njihovo skladenjsko vlogo v stavčni strukturi, in sicer jo najprej imenuje »strnitev, tj. zamena kakega stavka, npr.: *Na podstrešju je samo ena sobica -> … je ena sobica; drugih ni*« (Toporišič 1982: 333), kasneje pa termin nadomesti z izrazom stavčni skrčki: »Členki niso deli stavčnih zgradb, v okviru katerih se pojavljajo, ampak skrčki, ki nadomeščajo izpustne stavke, ki bi lahko ubesedovali sotvarje stavkov, v katerih so členki« (Toporišič 2000: 445). Pri tem navede podoben zgled kot pri strnitvi: *Sosedovi imajo* samo *enega otroka*, ki ga lahko razumemo kot *Sosedovi imajo enega otroka, imeli pa bi jih lahko več* (oz. *navadno je v družinah več kot en otrok*). Iz zgledov je razvidno, da gre za isto pojavnost. Zanimivo je, da Toporišič že od vsega začetka med členke uvršča tudi t. i. frazeološke členke oz. stalne besedne zveze, kot so npr. *kje neki*, *to se ve da ne*, *da le ne bi*, *tako rekoč*, *splošno govoreč*, *z drugimi besedami*, *kakor se reče*, *po pravici povedano*, *da po pravici povem*, *po mojem mnenju*, *prej ko ne*, *kako da ne*, *kako to da ne*, *tako tako*, *da da*, *nikakor ne* (gl. npr. Toporišič 1974: 277).

Avtor v prvi izdaji svoje slovnice (1976: 384–385) poda pomensko klasifikacijo členkov, ki je, kot sam pravi, narejena po slovaški akademijski slovnici (Toporišič 1982: 146). V tem smislu loči:

1. navezovalne členke, ki se navezujejo na predhodno izjavo (npr. *a*, *in*, *ja*, *zakaj potem*, *pa*, *potem pa*, *sicer pa*);
2. členke čustvovanja (npr. *hvala bogu*, *žal*, *dobro*);
3. poudarne členke (npr. *ravno*, *posebno*, *zlasti*, *predvsem*);
4. izvzemalne členke (npr. *le*, *samo*, *edino*, *komaj*);
5. presojevalne členke (npr. *kaki*, *blizu*, *približno*, *skoraj*);
6. dodajalne členke (npr. *tudi*, *prav tako*, *niti*, *poleg tega*);
7. členke zadržka (npr. *pravzaprav*, *pač*, *saj*, *komaj*, *že*, *sicer*, *resda*);
8. členke potrjevanja ali soglašanja (npr. *da*, *dejansko*, *gotovo*, *ja*, *kajpak*, *pravilno*, *prosim*, *res*, *seve*, *seveda*, *očitno*, *pri moji duši*);
9. členke možnosti ali verjetnosti (npr. *morda*, *mogoče*, *nemara*, *konec koncev*, *bržčas*, *bržkone*, *verjetno*, *komaj*);
10. členke mnenja ali domneve (npr. *baje*, *menda*, *češ da*, *denimo*, *recimo*);
11. vprašalne členke (npr. *ali*, *a*, *kaj*, *mar*, *kaj*, *kajne*);
12. spodbujalne členke, in sicer trdilne (npr. *a*, *da*, *kako ne*, *ko*, *morda*, *naj bo*) in nikalne (npr. *da*, *kaj*, *kakopak*, *ko*, *nikar*);
13. členke zanikanja in nesoglašanja (npr. *ne*, *nikar*, *ni*, *vraga*, *figo*).

Kasneje to klasifikacijo nekoliko modificira, v četrti izdaji slovnice pa ji doda novo, ki jo imenuje Prva klasifikacija (Toporišič 2000: 445–448). V novi klasifikaciji loči štiri osnovne tipe členkov:

1. pozivni (apelni), kjer se tvorec besedila obrača na naslovnika in na njegovo razmerje do sporočane resničnosti; v to skupino sodijo naslednji podtipi členkov:
   a) vprašalni (npr. *ali*, *ali ne*, *kaj ne*, *a*, *mar*),
   b) zahtevalni ali ukazovalni (npr. *naj*, *da*, *no*),
   c) želelni (npr. *naj*, *ko*, *da*),
   d) prepričevalni (npr. *res*, *resnično*, *dejansko*, *saj*, *gotovo*),
   e) zagotavljalni (npr. *dejansko*, *pri moji veri*, *gotovo*),
   f) nasprotovalni (npr. *toda*, *ko pa*, *vendar*, *saj*),
   g) grozilni (npr. *saj*, *boš že*, *še*),
   h) očitalni (*pa*);
2. vrednotenjski, ki podajajo tvorčevo razmerje do vsebine besedila:
   a) naklonski: gotovostni (npr. gotovo – *gotovo*, *očitno*, *nedvomno*, *tako in tako*; verjetno – *verjetno*, *pač*, *najverjetneje*; ni gotovo – *morda*, *nemara*, *navsezadnje*, *morebiti*, *konec koncev*, *bržčas*; skoraj gotovo – *komaj*, *težko*; ni gotovo – *bog vedi*, *vrag ve*, *kdo ve*, *baje*); hotenjski (npr. *nujno*),
   b) vrednotenjski primerjalni (npr. *dobesedno*, *tako rekoč*, *kakor*),
   c) popravni (npr. *torej*, *bolje*),
   d) domnevalni: nečustvenostni (npr. *kajpada*, *razumljivo*, *očitno*); čustvenostni (npr. *za čuda*, *toda*, *tako in tako*),
   e) merilni: nepolnomerilni (npr. *samo*, *prej*, *skoraj*); polnomerilni (npr. *dobesedno*); približne mere (npr. *okrog*, *okoli*, *blizu*),
   f) poudarjalni (npr. *sploh*, *še*, *niti za trenutek*, *edino*, *tudi*, *že*);
3. čustvenostni, ki podajajo čustvena razmerja, kot so presenečenje, čudenje, pomilovanje, pomiritev (npr. *hvala bogu*, *k sreči*, *boglonaj*), razočaranje (npr. *žal*, *na nesrečo*), malomarnost, bojazen (npr. *bog ne daj*);
4. besedilnozgradbeni, ki kažejo razmerje besedovalca do zgradbe besedila, in sicer:
   a) naznačujejo začetek besedila ali njegovega dela (npr. *nu*, *no*, *ja*, *torej*, *tako*),

b) členijo besedilo izrecno: premočrtno, in sicer naštevalno (npr. *obenem*, *poleg tega*, *nadalje*, *končno*, *prvič – drugič – tretjič*, *niti – niti*, *tako – kakor*) oz. opozarjalno (npr. *skratka*, *preprosto*, *nasprotno*, *slučajno*, *pravzaprav*, *potemtakem*, *natančneje*),

c) besedilo členijo sovsebno (implikativno), in sicer s poudarjanjem nasprotja (npr. *samo*) oz. z izbiro izrazov istega reda (npr. *na primer*, *recimo*, *čeprav*).

Kot je razvidno iz zgornjih tipov, je nova klasifikacija precej bliže funkcijskemu vidiku, vendar nekoliko preseneča, da pravzaprav nikjer ne poda teoretičnih izhodišč zanjo, zato je morda tudi nekoliko težje razumljiva. Vsekakor pa, kot pravi Smolej (2009: 15), »kaže v smer obveznega razumevanja členkov kot komunikacijske (povezovalne) in besedilne besedne vrste, kar pomeni, da naj bi bila osnovna vloga členkov izražanje razmerja govorečega do vsebine/sogovorca ali členjenja besedila«, s tem pa je nakazan tudi »možni novi pristop (besedilni, pragmatični) k razumevanju členkov«. In glede na to, da členki nimajo stvarnega pomena, je tudi delitev glede na pomen, kot pravi Černelič Kozlevčar (1993: 225), drugotna; precej bolj pomembna je namreč njihova funkcija v stavku, povedi in besedilu.

2.2 Tipološka klasifikacija členkov glede na funkcijsko-pomenski kriterij

Černelič Kozlevčar (1993: 225) meni, da je osnovno merilo za delitev členkov njihova vloga v stavku. V tem smislu loči členke, ki:

1. opravljajo določeno vlogo v stavku glede na skladenjski naklon: spodbujalni in vprašalni členki;
2. izražajo določeno razmerje do stvarnosti oz. resničnosti povedanega: možnostni oz. verjetnostni členki, členki mnenja, domneve, zadržka;
3. se navezujejo na posamezni stavčni člen, pri čemer člen v stavku izpostavljajo: poudarni, izvzemalni in presojevalni členki;
4. ne vežejo stavkov, ampak izražajo razmerje do sobesedila: navezovalni členki.

Kot je razvidno iz zgoraj navedenega, so podtipi nekoliko modificirana Toporišičeva klasifikacija.

Glede na funkcijo členkov v stavku, povedi oz. besedilu avtorica loči štiri osnovne podskupine:

1. skladenjskonaklonske členke: vprašalni (npr. *ali*), želelni (npr. *naj*);
2. naklonske členke v ožjem smislu, s katerimi govorec izraža svoje razmerje do vsebine ali do naslovnika;[1]
3. poudarne členke, ki kot skladenjski del stavčnega člena okrepijo stavek oz. poved;
4. navezovalne členke, ki vzpostavljajo pomenska razmerja s sobesedilom.

Funkcijsko-pomensko delitev členkov na dve temeljni skupini, na povezovalne in naklonske členke, prevzame tudi Žele (2015: 17) in, po njenem mnenju, takšna delitev ustreza tudi sporočanjsko-pragmatičnemu vidiku. V okviru vsake skupine loči naslednje podskupine:

1. povezovalni (besedilni) členki, ki izhajajo iz pragmatičnih okoliščin in poudarjajo besedilno koherentnost in kohezivnost:
   a) dodajalni (npr. *celo*, *kaj šele*, *še več*),
   b) izbirni (npr. *drugače*, *sicer pa*),
   c) izvzemalni (npr. *edino*, *le*, *sicer*),
   d) navezovalni oz. nadaljevalni (npr. *kakorkoli že*, *najsibodi*, *potemtakem*, *vendarle*, *vsekakor*),
   e) nadomestni (npr. *namesto tega*, *nasprotno*),
   f) nasprotovalni (npr. *zato pa*),
   g) pojasnjevalni (npr. *to se pravi*, *torej*),
   h) ponazarjalni (npr. *namreč*),
   i) popravni (npr. *ali bolje*, *namreč*, *oziroma*),
   j) poudarni (npr. *pravzaprav*, *predvsem*, *vsaj*, *zlasti*),
   k) povzemalni (npr. *skratka*, *torej*),
   l) zastranitveni (npr. *mimogrede*);
2. naklonski (medosebni) členki, ki izhajajo iz sporočanjskih razmerij in se osredotočajo na udeležence, okoliščine, glagolski proces ali količino:
   a) čustvenostni (npr. *bogvaruj*, *končno*, *saj*, *začuda*),
   b) pozivni (npr. *dejansko*, *kajne*, *naj*, *mar*),
   c) vrednotenjski (npr. *baje*, *morda*, *navsezadnje*, *nemara*, *takorekoč*, *verjetno*),
   d) členki zanikanja (npr. *nikar*).

1 Jakop (2000: 68) združi obe naklonski skupini v eno v širšem smislu.

V nadaljevanju doda, da členki, ne povezovalni ne naklonski, niso deli propozicije, temveč jo le modificirajo (Žele 2015: 17), o čemer podrobneje govori že Vidovič Muha (2013: 35), ko pravi, da je členek »vsaj v delu svojega obsega modificiran izraz globinskega stavčnega dela povedi, kar se površinsko kaže v njegovi vlogi modifikatorja stavka oz. katerega izmed stavčnih členov«. Členek je torej, kot pravi Žele (2014b: 322):

> »samostojna skladenjskofunkcijska besedna vrsta oz. skladenjskofunkcijski modifikator, ki pa nima niti predmetnega niti slovničnega samostojnega pomena niti ni oblikovno določena, in kot funkcijsko razpomenjena ali besedilno omejena besedna vrsta – saj jo sproti določa šele konkretna skladnja – je tudi stavčni nečlen oz. modifikator tipa *seveda*.«

Členki spreminjajo pomenska razmerja in vnašajo nova, z njimi se sporočevalec navezuje na kontekst, pri čemer izraža različne pomenske in čustvene odtenke posameznih izrazov ali povedi. In, po avtoričinem mnenju, se morajo vse te različne funkcije odražati tudi v slovarskih razlagah: »V ospredju je torej njihova vloga v besedilu oz. besedilna funkcija, zato so samo funkcijska beseda in slovarske razlage za členke so (samo) funkcijske; leksikografsko lahko členek opredelimo kot ubesedeno (slovarsko) referenco z govornim dejanjem« (Žele 2015: 17).

2.3 Tipološka klasifikacija členkov glede na besedilni oz. pragmatskosporočilni kriterij

Smolej (2004a: 142) meni, da sta potrebni in nujni tako semantična kot skladenjska raven preučevanja, vendar pa v analizah že ves čas »ostaja odprto in nerešeno vprašanje besedilotvorne funkcijske zmožnosti členkov in nadalje pragmatskosporočilne vrednosti, ki jo členki dobijo s konkretno rabo v povedi«. V tem smislu predlaga analizo členkov »le na ravni besedilnega okvirja oz. le na temelju sobesedila, v katerem so členki uporabljeni« (Smolej 2001: 48), pri čemer bi bila glavna kriterija za klasifikacijo njihovo skladenjsko vedenje in njihova zmožnost opravljanja besedilotvorne funkcije (ibid.: 96).

Smolej predstavi tri različne klasifikacije členkov.

Prva (Smolej 2001: 82) sledi funkcijsko-pomenskemu kriteriju, pri čemer, tako kot že njeni predhodniki, loči »dve osnovni nadskupini, ki pa se med seboj prepletata oz. ki sta med seboj povezani«, saj lahko nekateri členki (poudarjalni) opravljajo obe funkciji:

1. pomenski modifikatorji, ki omejujejo in natančno določajo pomen/cilj sporočilne funkcije, v povedi pa so vidni kot pomenska dopolnitev vsebine

oz. referencialnega pomena: modalni členki (npr. *morda*, *verjetno*, *gotovo*), poudarjalni členki (npr. *saj*, *če*, *da*), členki čustvovanja (npr. *žal*, *na srečo*), nikalni členki (kadar niso v funkciji odgovora na polno vprašanje);

2. besedilni povezovalci, ki so lahko a) sekundarni nosilci vsebine (anaforična sredstva), ki prenašajo vsebino predhodnih delov besedila oz. besedilnih polnopomenskih enot, b) napovedovalci modifikacije oblike oz. modifikacije leksikalnih sredstev, ki so most med dvema skupinama besedilnih polnopomenskih enot z isto ali podobno vsebino, c) delilni signali, ki spremljajo različne faze govornega dejanja (začetek, konec, njegovo prekinjanje, ponovno navezovanje ipd.).

Druga klasifikacija, ki členke razvršča glede na njihovo skladenjsko vedenje in njihovo zmožnost opravljanja besedilotvorne funkcije (Smolej 2001: 100–246):

1. pritrdilni in nikalni členki (npr. *da*, *ja*, *ne*, *nikar*, *nikakor*, *niti*, *nič*, *kje pa*);
2. tvorci stalnih sporočanjskih oblik povedi (npr. *a*, *ali*, *če*, *da*, *kaj*, *ko*, *mar*, *naj*, *saj*), kamor prišteva štiri osnovne vrste sporočanjskih dejanj (obvestilo, vprašanje, ukaz, želja) in še druge vrste (nasvet, zavrnitev, opozorilo, očitek, grožnja, graja, predlog itd.);
3. modalni členki:
   a) ki v povedi vplivajo na vrednost gotovostne naklonskosti (npr. *absolutno*, *brez dvoma*, *dejansko*, *gotovo*, *niti slučajno*, *očitno*, *prav*, *res*, *seveda*),
   b) ki ne vplivajo na vrednost gotovostne naklonskosti, ampak prenašajo subjektivno stališče govorca (npr. *baje*, *bržkone*, *menda*, *morda*, *najbrž*, *po vsej verjetnosti*),
   c) ki imajo obe funkciji;
4. členki čustvovanja (npr. *na srečo*, *žal*, *hvala bogu*);
5. poudarjalni členki (npr. *blizu*, *domala*, *izključno*, *malone*, *natančno*, *predvsem*, *ravno*, *še*, *šele*, *tudi*, *zgolj*, *zlasti*);
6. besedilni povezovalci:
   a) sekundarni nosilci vsebine (npr. *drugače povedano*),
   b) napovedovalci modifikacije oblike (npr. *kakor koli že*, *kljub vsemu*, *konec koncev*, *nenazadnje*, *se pravi*, *sicer*, *skratka*, *torej*, *vseeno*),
   c) delilni signali (npr. *prvič*, *drugič*, *dalje*, *za začetek*).

Tretja klasifikacija (Smolej 2001: 269) temelji na stopnji obveznosti členkov v strukturi, pri čemer se ločijo:

1. obvezni členki (členki pritrdilnosti, zanikanja, del tvorcev stalnih sporočanjskih oblik povedi, modalni členki z vlogo slabitve gotovostne naklonskosti, modalni členki z vlogo krepitve in slabitve gotovostne naklonskosti, kadar so odgovor na polno vprašanje, poudarjalni členki, kadar opravljajo tudi vlogo besedilnih povezovalcev, besedilni povezovalci):
   a) dvostopenjski (obveznost višje stopnje), pri čemer odstranitev členka poruši sporočanjsko obliko povedi in posledično tudi besedilno koherenco,
   b) enostopenjski (obveznost nižje stopnje), pri čemer odstranitev členka vpliva le na prekinitev besedilnega toka oz. besedilnega smisla;
2. neobvezni členek, ki ga lahko iz povedi odstranimo, ne da bi bila pri tem kakorkoli porušena sporočanjska oblika povedi ali zgradba besedila (poudarjalni nikalni členki, nekateri tvorci stalnih sporočanjskih oblik povedi, modalni členki z vlogo krepitve gotovostne naklonskosti, kadar niso odgovor na polno vprašanje, nekateri modalni členki, členki čustvovanja, poudarjalni členki, kadar ne opravljajo besedilotvorne vloge).

2.4 Tipološka klasifikacija členkov na osnovi sistemske funkcijske slovnice

Podobno kot predhodniki, tudi Skubic (1999: 211), ki izhaja iz Hallidayeve sistemske funkcijske slovnice (Halliday 1985), ugotavlja, da členki v stavčni strukturi ne morejo nastopati kot stavčni členi in se zdi, kot »da *stojijo* ob skupini besed (ki je lahko stavek ali katerikoli njegov del) in jo na neki način modificirajo, vplivajo na njeno vrednost v danem kontekstu«. Zato meni, da je namesto semantičnih meril precej pomembnejša njihova funkcija v strukturi in pomenski podstavi stavka. Za razliko od prislovov, ki imajo propozicijsko (okoliščinsko) funkcijo, imajo členki povezovalno oziroma modalno funkcijo (Skubic 1999: 229).

Klasifikacije členkov se Skubic loti iz povsem drugega zornega kota. Na primeru publicističnih besedil in z metodologijo sistemske funkcijske slovnice klasificira povezovalne členke, v klasifikacijo pa vključi tudi veznike, ki so sposobni izražati razmerje med povedmi, ter formalno in semantično integrirane strukture (prve so različni delni stavki, ki niso sestavni del propozicije, druge pa so sestavni deli propozicij). Pri tem loči tri tipe razširjanja besedila:

1. dodelava ali opis, kjer dodani stavek temeljiteje dodela prvega, in sicer ga ponovi z drugimi besedami kot:
   a) preoblikovanje, tj. ponovitev vsebine z drugimi besedami (veznik *in sicer*, členki *se pravi*, *z drugimi besedami*, *torej*, formalno integrirane strukture *to se pravi; da*, *lahko bi rekli; da*, *to pomeni, da*);

b) ponazoritev, tj. pojasnitev vsebine s konkretnimi podrobnostmi (vezniki *in sicer*, *in to*, členki *na primer*, *denimo*, *namreč*, formalno integrirane strukture *če ponazorim*; *za ilustracijo naj povem, da*),

c) pojasnilo, tj. spremenjena vsebina v obliki popravka (členki *namreč*, *ali bolje*, *oziroma*, formalno integrirane strukture *če smo natančnejši*), zastranitve (členki *mimogrede*, formalno integrirane strukture *ko/če smo že pri tem*), vrnitve (členki *torej*, formalno integrirane strukture *če se vrnem*, *če nadaljujem*, *kot sem rekel*), opustitve (vezniki *sicer pa*, členki *kakorkoli že*, *vsekakor*, *vendarle*, formalno integrirane strukture *kakorkoli pogledamo*, *če pustimo to ob strani*, semantično integrirane strukture *kljub vsemu*, *v vsakem primeru*), izpostavitve (členki *predvsem*, *še posebej*, *zlasti*, formalno integrirane strukture *posebej kaže izpostaviti, da*), povzetka (členki *skratka*, *na kratko*, *v glavnem*, formalno integrirane strukture *če povzamem*), podkrepitve (členki *pravzaprav*, *v bistvu*, *dejansko*);

2. podaljšanje, kjer novi stavek prejšnjega podaljša z novo vsebino:

   a) z dodajanjem nove informacije (vezniki *in*, *ne samo – tudi*, *pa*, *medtem pa*, členki *poleg tega*, *tudi*, *še več*, formalno integrirane strukture *naj dodam še to, da*),

   b) s protivnim dodajanjem, tj. vsebina je s prvo v protislovju (vezniki *pa*, *ampak*, *toda*, členki *po drugi strani*, *obenem/hkrati pa*, *zato pa*),

   c) s spremembo, ki je izražena kot nadomestitev (vezniki *pa*, členki *nasprotno*, *namesto tega*, *zato pa*), izvzemanje (vezniki *le da*, členki *drugače pa*, *sicer*, *razen tega*, formalno integrirane strukture *če to odmislimo*), zamenjava (vezniki *ali pa*, členki *lahko pa tudi*, *drugače pa*, formalno integrirane strukture *druga možnost bi bila, da*);

3. dopolnitev, v kateri novi stavek prejšnjega dopolni z okoliščinskimi informacijami, in sicer:

   a) časovno-prostorsko, ki je lahko prostorska (semantično integrirane strukture *tam*, *zadaj*, *pred njim*), časovna (vezniki *nakar*, *nato*, členki *končno*, *takoj*, semantično integrirane strukture *najprej*, *na začetku*, *poprej*, *obenem*, *naslednjič*), notranja oz. endoforična (členki *najprej*, *prvič*, *na tem mestu*, *drugič*, *in končno*, formalno integrirane strukture *če začnem*, *ob tem lahko omenim še, da*, *če nadaljujem*, *če sklenem*), načinovna (semantično integrirane strukture *na ta način*, *prav tako*, *drugače*),

   b) vzročno-pogojno, ki je lahko vzročno-posledična (vezniki *kajti*, *saj*, *zato*, členki *saj*, *namreč*, *navsezadnje*, *posledično*, formalno integrirane

strukture *iz tega sledi, da*, *se pravi, da*, *rezultat je, da*, semantično integrirane strukture *zaradi tega*, *ker je tako*, *glede na to*), pogojna (členki *sicer*, *drugače*, formalno integrirane strukture *kot stvari stojijo*, semantično integrirane strukture *v tem primeru*, *potem*, *če je tako*), dopustna (vezniki *a*, *ampak*, členki *resda*, *načeloma*, *vendar pa*, *po drugi strani pa*, semantično integrirane strukture *kljub temu*),

c) ozirno, tj. razlagati jo treba v tematskem okviru prejšnje povedi (semantično integrirane strukture *tukaj*, *glede tega*, *v zvezi s tem*).

3 POJMOVANJE ČLENKOV V SLOVENSKI LEKSIKOGRAFIJI

Neenotnost pri pojmovanju členka se odraža tudi v slovenski leksikografiji. V nadaljevanju sledi kratek prikaz obravnave členkov v različnih slovenskih jezikovnih priročnikih: Slovarju slovenskega knjižnega jezika (SSKJ1), Slovenskem pravopisu (SP), drugi izdaji Slovarja slovenskega knjižnega jezika (SSKJ2), Slovarju slovenskih členkov (SSČ), Slovenskem oblikoslovnem leksikonu Sloleks (Sloleks) in Presisovem večjezičnem slovarju (PVS). Vsi priročniki (razen SSKJ2) so prosto dostopni na spletni strani www.termania.net.

Glede na to, da je koncept za SSKJ1 (1970–1991) nastal v 60. letih prejšnjega stoletja, konceptualno temelji na predtoporišičevski slovnici, kar je razvidno tudi iz slovarske iztočnice *členek*, ki ima v lingvistiki pomen 'beseda brez samostojnega pomena, ki se navadno prideva drugim' oz., kot pravijo Bajec et al. (1956: 256), »členke imenujemo nepregibne besede, kadar ne živé več v samostojnem besednem pomenu, marveč so le sestavni del stalnih rekel ali besed, ki jim včasih dajejo poudarek ali pomensko tančino«, kot so npr. *le* (*tale*, *le-ta*), *li* (*ka-li*, *ali*), *si* (*lej si ga no*, *bodisi*), *ga* (*bog si ga vedi*), *bodi* (*kdorsibodi*), *koli* (*kjer koli*, *karkoli*, *kakorkoli*) itd.

Teoretične predpostavke o členku kot samostojni besedni vrsti, ki jih Toporišič opiše v svoji slovnici, se v praksi prvič uveljavijo v Slovarskem delu SP (2001), kot samostojna besedna vrsta pa so izpostavljeni tudi v prenovljenem SSKJ2, kjer se poleg členka pojavi novo kvalifikatorsko pojasnilo 'v členkovni rabi'. Nekoliko nenavadno pa je, da je vseh členkov v SSKJ2, kot pravi Krek (2014: 152), »193, a pogosto povsem drugih kot v SP2001«. Že v Uvodu je predstavljen tudi koncept slovarskih razlag: »Členek ima ali (nepolno) pomensko oziroma sinonimno (cirka, ipak, vešda, žalibog) ali pomensko-funkcijsko (ampak, edino, seveda, začuda) razlago« (SSKJ2: 32), iz česar lahko sklepamo, da slovarski opisi temeljijo predvsem na pomenskem opisovanju, ki pa se je izkazalo kot ne najbolj ustrezno.

Podoben pristop se nakazuje tudi v Osnutku koncepta novega razlagalnega slovarja slovenskega knjižnega jezika (Gliha Komac et al. 2015).

Skoraj istočasno kot SSKJ2 izide tudi SSČ (Žele 2014b), ki povzroči še dodatno zmešnjavo. Avtorica namreč v slovarju navaja členke v precej širšem smislu in, med drugim, kot členke označi tudi nekatere besede, ki so v SSKJ2 označene kot medmeti (npr. *ah*, *aha*, *eh*, *i*, *o*). Sama klasifikacija členkov v slovarju sicer temelji na teoretičnih predpostavkah, podanih npr. v Žele (2015), vendar pri podrobnejši analizi lahko vidimo, da teorija v praksi očitno ne deluje. Tako so npr. nekateri členki, ki so v klasifikaciji navedeni kot zgledi posameznih tipov, v slovarju razvrščeni v drugo podvrsto, npr. pri povezovalni členkih je nadomestni *nasprotno* razvrščen med nasprotovalne, pojasnjevalni *torej* med navezovalne, ponazarjalni *namreč* med pojasnjevalne itd., pri naklonskih členkih je čustvenostni *saj* uvrščen med vrednotenjske, pozivni *dejansko* med vrednotenjske, pozivni *mar* med čustvenostne itd. Nekateri členki pa so razvrščeni celo v drugo vrsto, npr. povezovalni dodajalni *kaj šele* med naklonske nikalne, pojasnjevalni navezovalni *kakorkoli že* med naklonske vrednotenjske, povezovalni navezovalni *vsekakor* med naklonske vrednotenjske itd.

Nekoliko drugačni kriteriji za določanje besednovrstnih kategorij so uporabljeni v priročnikih, nastalih po metodologiji korpusnega jezikoslovja, kot sta npr. Slovenski oblikoslovni leksikon Sloleks, ki je nastal v okviru projekta Sporazumevanje v slovenskem jeziku (2008–2013), ali Presisov večjezični slovar, narejen iz slovarja, ki ga uporablja strojni prevajalnik Presis. Oba sta namreč nastala na temelju Priporočil za oblikoslovno označevanje JOS. Avtorji (Erjavec et al. 2010) ugotavljajo, da so v jezikovnih priročnikih informacije o besednovrstni pripadnosti velikega deleža členkov zelo različne, saj so uvrščeni bodisi med členke bodisi med prislove, ponekod pa se kot dvojnice pojavljajo v obeh kategorijah. Klasificiranje besedne vrste glede na semantiko konteksta, tj. z uporabo vprašalnice, ki jo v primeru prislovov lahko zastavljamo, v primeru členkov pa ne, po njihovem mnenju ni ustrezno, ker se težave pojavljajo že na ravni ročnega določevanja, medtem ko je za avtomatsko analizo naloga težje izvedljiva (več o tem npr. Krek 2010). Ena od možnosti je, kot pravijo, »da se besedam, ki se v kontekstu lahko uporabljajo zgolj kot členki – to je, ki nikoli ne omogočajo zastavljanja vprašalnice – (ali pa so enakopisni z besednimi vrstami, ki omogočajo lahko razdvoumljanje, npr. s samostalniki), pripiše besednovrstna oznaka členek, vsem tistim, ki so enakopisni s prislovi, pa vedno le oznaka prislov«. Rezultati njihovih raziskav na korpusu FidaPLUS so pokazali, da je od 43 različnih besed 21 takih, ki so v leksikonu navedene le kot členek (*češ*, *ja*, *kajpak*, *komaj*, *le*, *menda*, *morda*, *najbrž*, *nemara*, *pač*, *pravzaprav*, *seveda*, *skoraj*, *sploh*, *še*, *šele*, *tudi*, *vsaj*, *zgolj*, *zopet*, *že*), dodatni štirje taki, ki so homonimni z lahko razdvoumljivimi besednimi vrstami (*celo*, *no*, *pa*, *torej*), ter 18 takih, ki so homonimni (tudi) s prislovi (*kar*, *končno*,

največ, *nazadnje*, *ne*, *okoli*, *okrog*, *prav*, *predvsem*, *približno*, *ravno*, *res*, *resda*, *resnično*, *samo*, *verjetno*, *zlasti*, *žal*), ki bi jih po novem označevali le kot prislove, kot členke pa ne. Podoznačevanje členkov v Priporočilih ni predvideno, predvsem zaradi nejasnih opredelitev te besedne vrste v strokovnih priročnikih.

4 ZAKLJUČEK OZ. KAKO NAPREJ

Kot je razvidno iz kratkega pregleda preučevanja členkov v slovenskem jezikoslovju, ostaja kljub različnim metodološko-raziskovalnim pristopom vprašanje členkov nerešeno. Tako se npr. začetni poskusi pomensko-skladenjske klasifikacije (gl. npr. Toporišič 1976), pri kateri se členek razvršča med slovnične oz. nepolnopomenske besedne vrste, ki za razliko od predmetnopomenskih oz. polnopomenskih vzpostavljajo slovnična razmerja med predmetnopomenskimi besedami, izkažejo kot neuspešni. Na drugi strani pa tudi funkcijske (gl. npr. Černelič 1991; Skubic 1999; Žele 2014) in besedilno-funkcijske (Smolej 2004) ne prinesejo jasnih rešitev. Dejstvo pa je, da vse raziskave potrjujejo, da lahko govorimo o členku kot o samostojni besedni vrsti oz., kot pravi Müller (2009: 18), je vsekakor »upravičeno in utemeljeno od prislova ločevati členek«. Čeprav so izhodišča raziskav različna, se kažejo v njih tudi nekatere skupne točke. V tem smislu bi lahko členek opredelili kot modifikacijsko besedno vrsto, ki doda pomensko vrednost jezikovni strukturi, na katero se nanaša. V dosedanjih raziskavah se nakazujeta dve temeljni podskupini členkov: 1) modalni oz. naklonski členki, ki jezikovno strukturo pomensko dopolnijo s tvorčevim odnosom do ubesedene stvarnosti, okoliščin, besedila ipd., in 2) povezovalni členki, ki povezujejo jezikovne enote na ravni besedila. Gorjanc (1998: 372) meni, da bi lahko členkovno skupino lahko definirali (a) z razvrstitvijo – nestabilni konektorski položaj, (b) z odsotnostjo kohezivne vezi, in (c) z modifikacijo dela besedila, ob katerega se razvrščajo. Predvsem to velja za povezovalno skupino. Členek v skladenjski strukturi nima skladenjske vloge oziroma, kot pravi Grošelj (2015), ne ustreza nobenemu od treh tradicionalnih besednovrstnih kriterijev, saj se ga kot samostojno besedno vrsto ne da opredeliti ne pomensko (prekriven z drugimi nepolnopomenskimi besednimi vrstami), ne oblikoslovno (prekriven z drugimi nepregibnimi besednimi vrstami) ne skladenjsko (nima stavčnočlenske vloge), zato bi bil pri njegovi nadaljnji klasifikaciji morda ustreznejši besedilni funkcijski pristop, ki se mu med navedenimi klasifikacijami še najbolj približa Skubic (1999), teoretično in praktično podprt s korpusnim pristopom, ki nam omogoča vzorčenje precej večje količine podatkov in obenem precej bolj sistematičen vpogled v jezikovne strukture. Pri tem se mi zdi, da homonimnost z drugimi besednimi vrstami, predvsem s prislovom, ne bi smela biti argument za neoznačevanje členka in se strinjam s Krekom (2010: 222), ko pravi, da se »najboljši kompromis zdi ta, da se množico členkovnih elementov razmeroma natančno določi, s čimer pri analizi potem vemo, kje lahko pričakujemo razlikovanje med členkom in drugimi elementi.«

Samo klasifikacijo pa bi bilo treba izvesti na podlagi rezultatov korpusne analize in ne obratno. Dosedanja praksa klasificiranja členkov na temelju jezikovnih občutkov, drugih klasifikacij ipd. in iskanja potrditev v praksi namreč ni obrodila sadov. Poseben premislek zahtevajo večdelni členki oz. členkovni frazemi, predvsem zato, ker so v nekaterih klasifikacijah poleg večdelnih členkov, kot so npr. *tako ali tako*, *kakorkoli že*, *vsekakor pa* ipd., kot členki navedeni tudi jezikovni elementi, ki sodijo k besedilnim kategorijam metabesedila oz. metadiskurza, kot je npr. *po mojem mnenju*, *z drugimi besedami*, *se pravi*, *po drugi strani*. Pokazalo se je tudi, da pomenske razlage pri opisovanju členkov v slovarjih niso ustrezne, saj so, kot pravi Müller (2009: 105), precej oddaljene od prave uporabnosti. Zato predlaga vključevanje situativne semantike v samo razlago ali, z drugimi besedami, pojasnilo, v katerih situacijah uporabimo določen členek, npr. pri členku *da* 'da rečem, kadar soglašam s sogovornikom ali s povedanim'. Takšni opisi so tudi za uporabnike slovarja precej bolj razumljivi, saj, kot je navedeno tudi v Predlogu za izdelavo Slovarja sodobnega slovenskega jezika, »eksplicitno navajanje slovničnih podatkov v slovarju pogosto zaradi prezahtevnega metajezika ne doseže uporabnikov oz. jih uporabniki največkrat ignorirajo« (Krek et al. 2013b).

Seveda je to le ena od možnosti opredelitve, klasifikacije in opisovanja členkov, zagotovo pa je izdelava novega slovarja slovenskega jezika izvrstna priložnost za premislek tudi o tej temi.

X
Moč množic in sodobno leksikografsko delo

Potencial množičenja v sodobni leksikografiji

Darja Fišer in Jaka Čibej

Abstract
Owing to increasing volumes of linguistic data and time constraints, the nature of lexicographic work has changed significantly in the past two decades. A number of steps in the dictionary production process have already been automated, but the algorithms developed are still far from perfect. Dictionary construction therefore still involves a number of routine but time-consuming and expensive manual procedures, for which experienced lexicographers are overqualified. This is why contemporary lexicography has started to explore options such as crowdsourcing, which save both time and financial resources without reducing the quality of the results by taking into account the key principles of microtask design and campaign management. This allows lexicographers to devote all their energy to expert work on the dictionary. This paper provides an overview of language resources that were successfully crowdsourced, the main aspects and characteristics of crowdsourcing, and the quality control mechanisms that ensure the success of this innovative method, which, if implemented correctly, could have a lasting impact on the overall workflow of lexicographic projects as well as the use and life cycle of lexicographic products.

Keywords: crowdsourcing, microtask design, crowd motivation, quality control, legal and ethical aspects of crowdsourcing

Ključne besede: množičenje, oblikovanje mikronalog, motivacija množičnikov, zagotavljanje kakovosti množičenja, pravno-etični vidiki množičenja

1 UVOD IN OPREDELITEV KLJUČNIH POJMOV

V zadnjem desetletju so se z razmahom spleta in ob vse večji meri digitalizacije dela pojavile številne oblike spletnega sodelovanja, pri katerih uporabniki na različne načine prispevajo k uresničitvi skupnega projekta. Poleg odprtokodnih projektov (npr. Linux) in kolaborativnih iniciativ (npr. Wikipedija) med nove oblike dela spada tudi množičenje[1] (angl. *crowdsourcing*), ki označuje postopek, pri katerem skupina ljudi (množica, angl. *crowd*) prispeva k doseganju določenega cilja, in sicer tako, da se delo razdeli med posameznike, od katerih vsak opravi manjši, obvladljiv del, ki ne zahteva veliko truda in časa, vsi prispevani deli, združeni v celoto, pa predstavljajo znaten dosežek (Howe 2008). Pri tem je pomembno izpostaviti, da množice ne sestavljajo nujno strokovnjaki z določenega področja, saj so številni projekti z uporabo množičenja pokazali, da ob ustrezni podpori in pripravi nalog tudi laiki zmorejo opravljati naloge, ki so bili doslej v izključni domeni ekspertov. Zaradi sodobne tehnologije in globalne razširjenosti spleta postaja izkoriščanje potenciala množičenja vse preprostejše, ugodnejše in učinkovitejše, o čemer pričajo uspehi številnih podjetij, ki uporabljajo moč množic za reševanje problemov, izboljšanje storitev ali ustvarjanje izdelkov (npr. Threadless, iStockPhoto in Procter & Gamble).

Angleški izraz crowdsourcing je vpeljal Jeff Howe leta 2006. Ker gre za relativno nov koncept dela, ki se pojavlja v več različicah in po mnenju nekaterih zajema tudi več oblik uporabniškega prispevanja,[2] še vedno ni povsem enotno definiran. Estellés-Arolas in González-Ladrón-de-Guevara (2012) izpostavita, da se različne definicije množičenja razlikujejo predvsem v obsegu dela, ki ga definirajo kot množičenje – nekatere namreč zajemajo praktično vso spletno sodelovanje, tudi soustvarjanje (angl. *co-creation*) in uporabniške inovacije (angl. *user innovation*). Zato z luščenjem definicij iz relevantne literature izdelata enotno definicijo, ki učinkovito loči množičenje od ostalih dejavnosti:

> Množičenje je vrsta spletnega sodelovanja, pri katerem posameznik, institucija, neprofitna organizacija ali podjetje s pomočjo javnega razpisa ali vabila skupini posameznikov z različnimi znanji, velikostjo in stopnjo heterogenosti predlaga prostovoljno opravljanje določene naloge. Opravljanje teh nalog, ki se lahko razlikujejo po težavnosti in načinu dela in pri katerih množica sodeluje z delom, denarjem, znanjem in/ali izkušnjami, vedno prinaša korist obema stranema. Uporabnik bo z opravljanjem naloge potešil določeno potrebo, npr. po zaslužku, družbenem odobravanju, dvigu samozavesti ali razvoju svojih sposobnosti, pobudnik množičenja pa bo lahko pridobil rezultat dela in ga izkoristil (ibid.: 9–10) [prevod: J. Č.].

1 V prispevku za angleški izraz crowdsourcing uporabljamo izraz množičenje, ki ponuja tudi številne druge izpeljanke, npr. množičiti (angl. to *crowdsource*), množičnik/množičnica (angl. *crowdsourcer*), množičeni projekti (angl. *crowdsourced projects*).

2 V angleščini obstajata npr. tudi izraza *crowdfunding* (množično financiranje) in *crowdvoting* (množično glasovanje), ki ju nekateri obravnavajo kot podpomenki množičenja. Obenem nekateri avtorji pojem *crowdsourcing* obravnavajo kot nadpomenko za vse vrste spletnega sodelovanja.

Vsaka vrsta spletnega sodelovanja torej ne spada nujno v kategorijo množičenja, za katerega je ključno, da je izpolnjenih več kriterijev: prisoten mora biti pobudnik (podjetje, organizacija ali posameznik), ki posameznike povabi k opravljanju določene (mikro)naloge, pri čemer imajo od tega korist tako posamezniki, ki so deležni bodisi finančne nagrade bodisi spodbude v drugi (lahko tudi nematerialni) obliki, kot seveda tudi pobudnik, ki lahko rezultat množičenja uporabi pri nadaljnjem delu.

Na področju leksikografije je od omenjenih oblik uporabniškega prispevanja najbolj razširjena kolaborativna leksikografija, pod katero spadajo znani projekti, kot so Wikislovar,[3] Urban Dictionary,[4] Folkets lexikon[5] in slovenski Razvezani jezik.[6] Uporabniški prispevki h gradnji slovarjev so danes večinoma omejeni na kolaborativne leksikografski projekte, kot so prispevanje morebitnih slovarskih iztočnic in primerov ali pa na popravljanje slovarja po izdaji, ugotavljata Abel in Meyer (2013). Hkrati pa so se okoliščine v sodobni leksikografiji tako spremenile, da se leksikografi pri svojem delu soočajo z vse večjimi časovnimi omejitvami in količinami podatkov, zato je vse več leksikografskih postopkov (pol)avtomatskih. Nekatere stopnje gradnje slovarjev se zato spreminjajo v rutinska opravila, za katera so leksikografi prekvalificirani. Ob tem se ponuja priložnost, da v leksikografsko delo vključimo uporabniške prispevke v obliki množičenja – ne kot glavno fazo izgradnje slovarja, temveč kot način za filtriranje, obdelavo in čiščenje (avtomatsko luščenih) podatkov, ki nato leksikografom omogočijo hitrejšo izdelavo geselskega članka.

Čeprav je zadnjih nekaj let v leksikografskih razpravah vse več govora o množičenju, metoda še ni bila zadostno preizkušena na obsežnih, raznolikih leksikografskih projektih. V prispevku zato predstavljamo domače in tuje primere dobrih praks, različne vidike uporabe množičenja in načine, kako zagotoviti uspešnost te inovativne metode, ki bi ob ustrezni implementaciji lahko trajno vplivala na izvedbo leksikografskih projektov ter na rabo in življenjsko dobo leksikografskih izdelkov.

2 PREGLED UPORABE MNOŽIČENJA ZA PRIDOBIVANJE JEZIKOVNIH PODATKOV

V tem razdelku predstavljamo pregled sorodnih raziskav in projektov z različnih področij obdelave naravnih jezikov, ki so uspešno uporabili množičenje.

3 http://sl.wiktionary.org/ (dostop 8. 8. 2015).
4 http://www.urbandictionary.com/ (dostop 8. 8. 2015).
5 http://folkets-lexikon.csc.kth.se/ (dostop 8. 8. 2015).
6 http://razvezanijezik.org/ (dostop 8. 8. 2015).

2.1 Množičenje pri izdelavi in označevanju jezikovnih virov

Klubička in Ljubešić (2014) sta s pomočjo množičenja izdelala oblikoskladenjsko označen in lematiziran korpus hrvaščine, ki ga bo pozneje mogoče uporabiti kot učno množico. Evalvacija množičenja je pokazala, da je bila natančnost posameznega množičnika v povprečju 90 %, natančnost večinskega glasu treh množičnikov pa približno 97 %.

Venhuizen et al. (2013) so s pomočjo spletne aplikacije Wordrobe[7] množičnikom v reševanje ponudili naloge za določanje besedne vrste prekrivnih besednih oblik, kategorizacijo lastnih imen in označevanje pomena večpomenskih besed za razvoj in jezikoslovno označevanje angleškega korpusa Groningen Meaning Bank.[8] Rezultati množičenja so že ob majhnem številu množičenih podatkov dosegli visoko natančnost glede na zlati standard.

Rumshisky (2011) in Rumshisky et al. (2012) so z množičenjem na platformi Amazon Mechanical Turk pridobili pomensko označen korpus in pomenski leksikon za angleščino, ki so ju označili materni govorci – nestrokovnjaki. Rezultati kažejo, da tudi tak izdelek dosega kakovostni nivo, kakršnega bi dosegli strokovnjaki, označevanje pa je obenem zelo hitro in ekonomično, saj je kompleksno označevanje razdeljeno na več preprostejših korakov, ki jih zmorejo tudi naključni uporabniki platforme Amazon Mechanical Turk.

Fossati et al. (2013) so platformo CrowdFlower uporabili za označevanje semantičnih vlog v angleških besedilih. Metodo so primerjali s standardno metodologijo označevanja in ugotovili, da množičenje, ki temelji na označevanju po več (enostavnejših) korakih, dosega boljše rezultate kot standardno označevanje, saj je natančnejše in obenem tudi hitrejše.

2.2 Množičenje za jezikovne tehnologije

Eden najbolj tipičnih jezikovnotehnoloških nalog, ki jih znanstveniki razrešujejo s pomočjo množičenja, je strojno prevajanje, pri čemer množico uporabljajo za različne namene. Zaidan in Callison-Burch (2011) množičnike najemata za izbiro najboljšega strojnega prevoda med več ponujenimi kandidati in v eksperimentih pokažeta, da množičniki dosegajo kvaliteto, ki je primerljiva s profesionalnimi prevajalci. Z množičenjem je mogoče uspešno opraviti tudi evalvacijo strojnih prevajalnikov (Bentivogli et al. 2011; Denkowski in Lavie 2010), izvajati

7 http://wordrobe.housing.rug.nl/ (dostop 8. 8. 2015).
8 http://gmb.let.rug.nl/ (dostop 8. 8. 2015).

zanesljivo besedno poravnavo vzporednih korpusov (Gao in Vogel 2010) in izdelati kvalitetne učne korpuse za statistične strojne prevajalnike (Negri et al. 2011).

Chamberlain et al. (2008) z množičenjem uspešno zbirajo podatke za razreševanje sklicev v angleščini, in sicer s pomočjo spletne igre Phrase Detectives,[9] pri kateri množičniki označujejo besede ali besedne zveze, na katere se nanašajo zaimenske oblike. Na voljo je tudi različica na družbenem omrežju Facebook.

Platformo za množičenje Amazon Mechanical Turk so uporabili tudi Snow et al. (2008), med drugim za analizo sentimenta v naslovih angleških časopisov. Pri ocenjevanju podatkov, ki so jih označili nestrokovnjaki, so ugotovili, da je za vsako nalogo potrebno le majhno število odgovorov, da dosežejo enak rezultat, kot če bi jo reševal strokovnjak. V povprečju so za enako kakovost kot z eno oznako strokovnjaka potrebovali štiri oznake nestrokovnjakov.

2.3 Množičenje za slovenščino

Množičenje je bilo že uspešno uporabljeno tudi za izdelavo jezikovnih virov za slovenščino.

Fišer et al. (2014) so posebej razvito orodje za množičenje sloWCrowd (Tavčar et al. 2012) uporabili za odpravljanje napak v avtomatsko zgrajenem semantičnem leksikonu sloWNet. Množičniki so z orodjem glasovali, ali avtomatsko generirani kandidati sodijo v določeno množico sinonimov ali ne. Eksperiment je pokazal, da je sloWCrowd enostaven za uporabo tako za administratorje kot tudi za množičnike. Zbrani odgovori so glede na visoko stopnjo ujemanja med množičniki zanesljivi (povprečna natančnost dosega dobrih 80 %, kar je za kompleksne semantične naloge visoka vrednost), število dvoumnih rešitev pa zelo majhno.

Množičenje so za čiščenje avtomatsko luščenih kolokacij in zgledov iz korpusa Gigafida ter razvrščanje kolokacij in njihovih zgledov v (pod)pomene preizkusili tudi Kosem et al. (2013). Rezultati eksperimenta so pokazali, kako pomembna za doseganje zanesljivih rezultatov je jasna formulacija vprašanja, ki ne sme biti večdimenzionalno in subjektivno.

Na spletu je na voljo tudi Igra besed,[10] ki služi zbiranju kolokacij v slovenščini, v kateri igralci predlagajo po tri najbolj tipične pridevniške ali samostalniške kolokatorje za naključno izbran samostalnik ali pridevnik, za svoje odgovore pa prejemajo točke glede na seznam najmočnejših kolokacij, izluščen iz korpusa Gigafida (na podlagi

9 http://anawiki.essex.ac.uk/phrasedetectives/ (dostop 8. 8. 2015).
10 http://www.igra-besed.si (dostop 8. 8. 2015).

besednih skic v konkordančniku Sketch Engine). Igra ponuja način za enega igralca, igro z izbranim soigralcem ter igro z naključnim soigralcem, ki je ob istem času prijavljen na strežnik. Z igro se zbirajo podatki o besedah, ki jih igralci vpišejo, pa tudi kdo (uporabniško ime), kdaj in s kom igra. Ti podatki bodo uporabljeni za primerjavo s seznami kolokacij, da bo mogoče preveriti, katera kolokacijska mera najbolje zazna jezikovni čut ljudi. Rezultati imajo velik potencial tudi za čiščenje avtomatsko luščenih podatkov, kar pa bi bilo treba še analizirati.

3 MOTIVACIJSKI VIDIKI MNOŽIČENJA

Proces množičenja vedno vključuje pobudnika, ki od množice pričakuje opravljanje določene naloge, in množico posameznikov, ki v zameno za delo prejmejo plačilo oziroma nadomestilo. Kot pišeta Estellés-Arolas in González-Ladrón-de-Guevara (2012), plačilo oz. nadomestilo služi kot motivacija za množičnika, da delo opravi oziroma nadaljuje z njim.

Motivacija je torej pri množičenjskih projektih ključnega pomena, zlasti v primeru manjših jezikov, ki nimajo na voljo velike baze množičnikov in je zato treba tiste, ki so na voljo, še dodatno motivirati za delo. Motivacija je lahko bodisi materialna bodisi nematerialna, vedno pa mora izpolniti eno ali več potreb množičnikov, kot so gmotna nagrada, družbena prepoznavnost, dvig samozavesti, razvoj posameznikovih sposobnosti. Nagrado zagotovi pobudnik množičenja kot poplačilo za delo množice. Lew (2013) v razpravi o motivaciji uporabnikov za dodajanje uporabniških vsebin na splet loči tri kategorije motivacije (psihološko, družbeno in ekonomsko), ki veljajo tudi za množičenje, zato jih podrobneje predstavljamo v nadaljevanju.

3.1 Družbena motivacija

Družbeni vidik motivacije črpa iz potrebe posameznikov, da se povezujejo z drugimi posamezniki, ki imajo podobne interese, s sodelovanjem pridobivajo nova znanja ali spretnosti in povečujejo svoj ugled v skupnosti.

3.1.1 Pripadnost skupnosti

Večjo vlogo kot velikost skupnosti igra angažiranost njenih članov. Pomembno je torej, da se člani poistovetijo s skupnostjo in v njej navdušeno delujejo, ker želijo prispevati k njenemu uspehu, razvoju ali prepoznavnosti. Množičenje se v tem

primeru zanaša na pripravljenost posameznikov, da prispevajo k projektu, ki je v interesu in v korist vsej skupnosti.

Večina projektov kolaborativne leksikografije temelji prav na družbeni motivaciji, npr. že omenjeni Wikislovar, Urban Dictionary in Razvezani jezik, ki se je za slovenščino izkazal za uspešnega: v 10 letih trajanja projekta je okoli 1.600 anonimnih piscev prispevalo več kot 3.700 gesel in 2.300 člankov (Dolar 2014). To dokazuje, da so tudi v Sloveniji posamezniki pripravljeni sodelovati pri zbiranju leksikografskih podatkov. Da bi bili slovenski uporabniki družbenih omrežij v ustreznih okoliščinah pripravljeni prispevati tudi k izgradnji jezikovnih virov za slovenščino, lahko sklepamo iz aktivnih in konstruktivnih uporabniških skupin z jezikovno tematiko na Facebooku, npr. Prevajalci, na pomoč!, Za vsaj približno pravilno uporabo slovenščine, Skupina za ohranjanje rodilnika pri zanikanju, Društvo ljubiteljskih pravopisarjev in slovničarjev, Razgibane vejice ipd.

3.1.2 Izobraževalna motivacija

Posebna podvrsta družbene motivacije je izobraževalna, pri kateri se množičniki z reševanjem nalog učijo določene vsebine ali spretnosti. Ustrezno pripravljene naloge bi torej lahko ponudili v reševanje v sklopu rednih učnih vsebin ali kot dodatno gradivo za vaje na različnih izobraževalnih stopnjah. Tovrstni motivacijski pristop ima npr. spletna stran za učenje jezikov Duolingo[11] (von Ahn 2013), ki uporabnikom ponuja brezplačne tečaje tujih jezikov. Tečaji sestojijo iz različnih nalog, med drugim tudi iz stavkov, ki jih uporabniki za vajo prevajajo v tuji jezik in hkrati pripomorejo k prevajanju spletnih vsebin v druge jezike.

3.1.3 Priznanja in nazivi

Pod družbeno motivacijo spadajo tudi priznanja, ki jih množičnik prejme kot nagrado za delo v skupnosti. Lahko gre za priznanje v fizični obliki (npr. potrdilo o sodelovanju), prestižen naziv (npr. urednik Wikipedije) ali navedbo v dvorani slavnih projekta oz. skupnosti (angl. *hall of fame*).

3.2 Psihološka motivacija

Za mnoge uporabnike je dodajanje vsebin na splet psihološko izpolnjujoče, npr. ker radi delijo znanje z drugimi, ker tako izpolnjujejo potrebo po tem, da

11 https://www.duolingo.com/ (dostop 8. 8. 2015).

izražajo sami sebe, ali ker se jim sodelovanje zdi zabavno. Vidik zabave je bil podlaga za osnovanje t. i. iger z namenom, ki v zadnjem času postajajo ena najpopularnejših oblik sodela in množičenja, zato se jim v nadaljevanju razdelka podrobneje posvečamo.

3.2.1 Igre z namenom

Igre z namenom (angl. *games with a purpose*) so igre, ki jih uporabniki primarno igrajo zaradi lastnega zadovoljstva, obenem pa z igranjem pomagajo pri zbiranju podatkov. Vse več ljudi ima dostop do spleta (veliko od teh igra tudi računalniške igre), nalog, ki jih računalniki ne zmorejo opraviti brez človeške pomoči, pa je kljub tehnološkemu napredku še vedno veliko. Kot piše von Ahn (2006), je igre z namenom zato mogoče uporabiti na različnih področjih, npr. za izboljšanje iskanja po internetu in za filtriranje vsebin, v številnih primerih pa je bil ta način zbiranja podatkov uporabljen tudi pri raziskavah z jezikovnimi podatki.

Uspešna primera iger z namenom sta ESP Game (von Ahn in Dabbish 2004) in Peekaboom (von Ahn 2006), s katerima je bilo dokazano, da lahko množica reši probleme, ki jih računalniki še ne zmorejo. Pri igri ESP Game se v paru znajdeta igralca, ki se med seboj še ne poznata, obema pa se prikaže slika. Cilj igre je uganiti, s katero besedo bo partner označil sliko. Igra je zelo uspešna, zato je bilo v kratkem času označeno veliko število slik, podatki pa so bili uporabni npr. za izboljšanje internetnih iskalnikov in za razvoj programske opreme za slabovidne. Na podoben način deluje tudi Peekaboom, le da igralci določajo, kje na sliki se nahaja določen predmet, podatki pa se na to uporabijo za strojno učenje računalniškega vida.

Med uspešnimi igrami z namenom so tudi JeuxDeMots (Joubert in Lafourcade 2012), igra za gradnjo leksikalne mreže francoščine; že omenjena igra Phrase Detectives (Chamberlain et al. 2008); Puzzle Racer (Jurgens in Navigli 2014), igra za označevanje slik s pomeni; in Verbosity (von Ahn et al. 2006), s katero so s pomočjo vprašanj ali dopolnjevanja stavkov zbirali splošno znane podatke (npr. izjave, kot je *mleko je belo*), ki so nato uporabni za gradnjo ontologij in izboljšanje inteligence računalniških sistemov.

Veliko število iger z namenom kaže, da je igrifikacija (angl. *gamification*, predstavitev oziroma oblikovanje orodij in aplikacij v obliki iger) danes pri zbiranju jezikoslovnih podatkov že precej pogosta praksa, ki prinaša dobre rezultate, uporabne na raznolikih področjih.

3.3 Ekonomska motivacija

Ekonomska motivacija temelji na denarnih plačilih za opravljanje nalog oziroma na drugih gmotnih nagradah.

3.3.1 Mikroplačila

Denarno nadomestilo je pogosto pri velikih (zlasti komercialnih) projektih, pri katerih se od množičnikov pričakuje, da opravijo večjo količino dela, v katerega so vključeni dalj časa. Denarna nadomestila se najpogosteje izplačujejo v obliki t. i. mikroplačil (angl. *micropayments*), ki jih množičnik prejme za opravljeno nalogo oziroma za vnaprej določeno število opravljenih nalog. Na tak način delujejo številne znane platforme za množičenje, med drugim tudi Amazon Mechanical Turk,[12] CrowdFlower[13] in Clickworker.[14]

Postopek množičenja z mikroplačili je naslednji: pobudnik množičenja na platformo naloži projekt (sveženj mikronalog) in lastniku platforme vnaprej nakaže določeno količino denarja (odvisno od velikosti projekta, števila različnih nalog, zahtevnosti nalog ipd.). Določen del zneska pripada lastniku platforme za gostovanje projekta, ostalo pa lastnik platforme razdeli med množičnike glede na opravljeno delo.

Mikroplačila preko platform za množičenje so za motivacijo množičnikov uporabili že številni avtorji jezikoslovnih raziskav (Akkaya et al. 2010; Rumshisky 2011; Rumshisky et al. 2012; Fossati et al. 2013), a je treba omeniti, da imajo platforme, kot je npr. Amazon Mechanical Turk, svoj nabor množičnikov (registriranih uporabnikov, ki lahko rešujejo naloge), večinoma pa gre za angleške govorce (oziroma govorce večjih jezikov). Registriranih govorcev manjših jezikov na tovrstnih platformah ni dovolj, zaradi lokalne finančne in davčne zakonodaje pa se lahko zaplete tudi pri ustvarjanju računov za izvajanje množičenjskih projektov in nakazovanju mikroplačil.

3.3.2 Ostale nagrade

Ekonomska motivacija vključuje tudi druga nadomestila, kot so npr. kuponi, vstopnice, licence za programsko opremo in druge predmetne nagrade (majice, priponke ipd.). Po plačilih v tej obliki najpogosteje posegajo manjši projekti z

12 https://www.mturk.com/ (dostop 8. 8. 2015).
13 http://www.crowdflower.com/ (dostop 8. 8. 2015).
14 http://www.clickworker.com/en/ (dostop 8. 8. 2015).

omejenim financiranjem. Primeri dobre prakse (El-Haj et al. 2014; Fišer et al. 2014) kažejo, da množičnike pritegnejo tudi tovrstne nagrade, pogosto v kombinaciji z družbeno in psihološko motivacijo.

4 PRAVNI, FINANČNI IN ETIČNI VIDIKI MNOŽIČENJA

V tem razdelku predstavljamo pravne, finančne in etične omejitve, na katere naletimo pri uporabi množičenja. Pravne in finančne omejitve so v veliki meri odvisne od lokalne zakonodaje in financiranja, in čeprav ne vplivajo neposredno na kakovost in vsebino projekta, pogosto predstavljajo veliko oviro pri uvedbi množičenja v raziskovalno delo, še zlasti na področju leksikografije. Večina raziskovalcev namreč ni seznanjena s pravnimi omejitvami na tem področju, pomoč pravnih strokovnjakov pa je redka. Ker je množičenje še vedno relativno nova oblika dela, ni izrecno predvideno v zakonodaji, zato marsikatero vprašanje ostaja odprto.

4.1 Plačevanje množičnikov

Kot pišejo Sabou et al. (2014), je etična dolžnost pobudnika množičenjskega projekta, da v primeru ekonomske motivacije množičnikov upošteva realne življenjske stroške in mikroplačila prilagodi tako, da zneski v povprečju presegajo lokalno minimalno plačo oziroma da odsevajo urno postavko, ustrezno za tovrsten način dela. Tudi Fort et al. (2014) opozarjajo, da množičenje kot nova oblika dela še vedno ni obravnavano v delovni zakonodaji, kar postavlja množičnike v kočljiv položaj z vidika višine plačila, varnosti pri delu, delavskih pravic ipd. Na platformi Amazon Mechanical Turk naj bi se tudi do 20 % delavcev preživljalo le z reševanjem množičenjskih nalog, zato je ključno, da se jim zagotovi ustrezen zaslužek. Silberman et al. (2010) izpostavijo tudi dejstvo, da pobudniki množičenjskih projektov pogosto zamujajo s plačili. Na platformi Amazon Mechanical Turk mora pobudnik odobriti, da je bilo delo ustrezno opravljeno, preden lahko množičnik prejme plačilo. Platforma naloge samodejno odobri po 30 dneh, če tega prej ne stori pobudnik, a to pomeni, da lahko množičnik čaka do konca izteka roka, nakar pobudnik njegovo delo zavrne, množičnik pa ne dobi plačila, ki ga je pričakoval. Tovrstnim praksam se je treba izogibati.

Sabou et al. (2014) priporočajo, da se pred množičenjem izvede pilotno reševanje nalog in predhodno določi, koliko časa naj bi delo trajalo. Nekatere mikronaloge so težje in bolj zapletene od drugih, zato od množičnika zahtevajo več truda in

časa. Pri takšnih nalogah je mikroplačilo ponavadi višje, da je tudi urna postavka primerljiva. Ta vidik že upoštevajo npr. Krek et al. (2013), ki za lažje množičenjske naloge predvidevajo mikroplačilo 0,02 € na odločitev (kar pri približno 350 odločitvah na uro znese 7 €), za težje naloge pa 0,05 € (odločitev na uro je v tem primeru nekoliko manj, postavka pa je podobna). Cena je vsekakor odvisna tudi od proračuna projekta in količine podatkov, ki jo je treba obdelati. Pri izplačevanju mikroplačil je treba upoštevati obstoječe načine plačevanja (v Sloveniji npr. avtorske pogodbe, plačilo preko s. p., plačilo po študentski napotnici) in morebitne omejitve v davčni zakonodaji.

Upoštevanje etičnih načel je še toliko pomembnejše, če bodo zbrani podatki uporabljeni v komercialne namene in bodo ponudniku projekta prinesli zaslužek. V takem primeru je sporno, da množičniki za delo ne prejmejo plačila ali da je plačilo nerazumno nizko.

4.2 Omejitve pri najemanju množičnikov

Pri izbiranju množičnikov za projekt je treba imeti v mislih, da morda pri tem obstajajo pravne omejitve. To še zlasti velja v primeru mladoletnih delavcev (npr. dijaki), pri katerih je treba pridobiti predhodno soglasje staršev.

4.3 Priznavanje avtorstva

Ker množičniki na projektu pogosto opravijo nezanemarljiv delež dela, je treba vnaprej določiti, kako in kje se jim pripiše zasluge (npr. ali so navedeni kot soavtorji). Čeprav za navedbo avtorstva v primerih množičenja ni na voljo jasnih smernic, nekateri prostovoljski projekti (npr. FoldIt,[15] Phylo[16]) množičnike navajajo na seznamu avtorjev.

4.4 Podpis soglasja in obveščanje množičnikov o projektu

Ponavadi množičniki pred začetkom dela podpišejo soglasje, s katerim jih pobudnik množičenja obvesti o naravi projekta in o namenih, za katere bodo podatki uporabljeni. Množičnikom mora biti jasno predstavljeno, da bodo podatki npr. uporabljeni v raziskovalne namene in ali bodo po koncu projekta dostopni tudi tretjim osebam (zlasti če gre za odprte licence, kot je Creative Commons).

15 https://fold.it/portal/ (dostop 8. 8. 2015).
16 http://phylo.cs.mcgill.ca/ (dostop 8. 8. 2015).

4.5 Dostopnost podatkovnih zbirk

V primeru, da bodo podatkovne zbirke, ki bodo z množičenjem nadgrajene, prosto dostopne, je treba zanje izbrati primerno licenco v skladu z lokalno zakonodajo o avtorskih pravicah in varstvu osebnih podatkov.

5 MIKRONALOGE

Osnovna ideja množičenja je, da obsežen in kompleksen problem razdeli na manjše, obvladljive in enostavnejše dele. Celoten sklop dejavnosti, potrebnih za reševanje zastavljenega problema, imenujemo množičenjska kampanja, posamezne dele, ki jih v reševanje dobivajo množičniki, pa mikronaloge. Oblikovanje mikronalog je ključna stopnja v kateremkoli množičenjskem projektu. Zato v tem razdelku predstavljamo načela, ki jih je treba upoštevati pri izdelavi mikronalog, če želimo z množičenjem doseči kakovostne in predvsem uporabne rezultate, in navedemo še nekaj primerov uspešno oblikovanih mikronalog.

5.1 Načela oblikovanja mikronalog

Nezahtevnost – Ker pri reševanju mikronalog pogosto sodelujejo nestrokovnjaki, je pomembno, da so naloge kognitivno kar se da nezahtevne. Reševanje ene same naloge od množičnika ne sme zahtevati pretiranega razmisleka, saj je bistveno, da v čim krajšem času reši čim več nalog (prim. Rumshisky 2011; Snow et al. 2008; Lease in Alonso 2014).

Ustrezna vprašanja – Mikronaloge naj ne vsebujejo vprašanj, ki pri množičenju ne bodo dala dobrih rezultatov. Izključiti je treba predvsem nejasna ali dvoumna vprašanja in prekomerno subjektivne in nezanesljive ocene, saj se rezultati, pridobljeni iz takšnih nalog, pogosto izkažejo za nezanesljive in neuporabne (prim. Kosem et al. 2013). Zastavljena vprašanja morajo biti enodimenzionalna, zato je v primerih, ko gre za kompleksen, večplasten problem, priporočljivo, da se naloga razdeli na več preprostejših korakov (prim. Biemann in Nygaard 2010).

Prilagojenost ciljni skupini – Različne mikronaloge lahko od množičnikov zahtevajo različno stopnjo predznanja. Uvajanje množičnikov v postopek označevanja mora biti čim krajše, zato je treba za vsak skupek mikronalog izbrati ustrezno ciljno skupino glede na potrebno predznanje (npr. nestrokovnjaki, študenti ali strokovnjaki). Množičniki z nezadostnim predznanjem potrebujejo več uvajanja (kar je časovno neugodno), dajali pa bodo manj zanesljive in posledično manj

uporabne rezultate. Po drugi strani pa je strokovnjake, ki morajo reševati trivialne naloge, težje motivirati, medtem ko njihovo delo prav tako zahteva višje plačilo.

Tehnična preprostost in uporabniku prijazen vmesnik – Reševanje mora biti nezahtevno tudi z logističnega vidika, npr. da zahteva čim manjše število klikov z miško, čim manj premikanja po zaslonu in, če je le mogoče, čim manj tipkanja in vnašanja podatkov. Množičniki naloge najpogosteje rešujejo s pomočjo platform za množičenje (Amazon Mechanical Turk, Clickworker ipd.). V primeru, da se za projekt razvija lastna platforma, je treba zagotoviti, da ima uporabniku prijazen vmesnik, ki bo omogočal neteţavno registracijo, tekoče reševanje nalog in prehajanje med njimi. Množičniku je treba zagotoviti tudi možnost, da primer označi kot nejasen oz. ga preskoči, če npr. iz danega konteksta ne more jasno sklepati, kako bi ga označil, ali kot nerešljiv, če npr. nobena od danih oznak ni ustrezna. Pri igrah z namenom von Ahn in Dabbish (2008) izpostavita, da se mora igra končati v kratkem času, Jurgens in Navigli (2014) pa poudarita, da je vmesnik ključnega pomena – prednost je, če igra sploh ne vsebuje jezikoslovne terminologije, kadar večino podatkov zbira množica nestrokovnjakov, saj mora igra biti zanje čimbolj razumljiva in preprosta.

Kratka navodila – Navodila za reševanje mikronalog morajo pojasniti namen množičenjske kampanje, morajo biti jasna in kratka, priporočljivo je tudi, da vsebujejo ponazoritev na primeru.

Povratna informacija – Priporočljivo je, da množičnik za svoje odgovore dobi povratno informacijo. Na tak način se lahko nauči nekaj novega, obenem pa ga pravilni odgovori motivirajo, da z delom nadaljuje, saj lahko sproti preverja razumevanje vprašanj in pravilnost odgovorov ter se postopoma izboljšuje v reševanju problema.

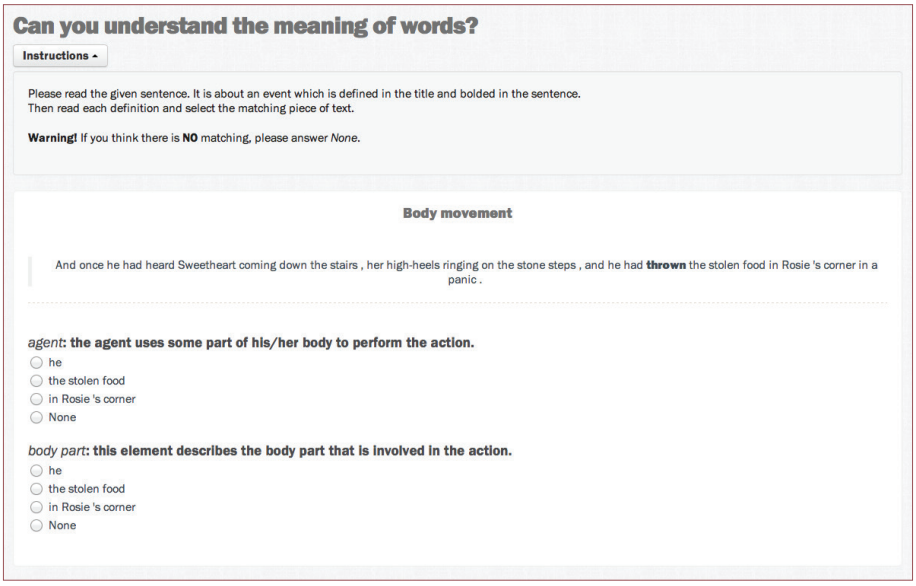
Izziv, naključnost in časovna omejitev – Nekaj vidikov, na katere je treba biti pozoren pri izdelavi iger z namenom, izpostavita tudi von Ahn in Dabbish (2008), ki so relevantne za vse množičenjske kampanje. Bolj kot je naloga zabavna, bolj je učinkovita. Zabavnost naloge zagotovimo tako, da jo zasnujemo kot izziv za igralca, npr. z uvedbo točkovnega sistema, časovne omejitve za reševanje naloge, lestvic množičnikov z najvišjim številom doseženih točk ipd. Ključno je tudi, da je število nalog, ki jih mora množičnik opraviti v določenem časovnem obdobju, tempirano tako, da mu je v izziv (da torej naloga ni niti preveč preprosta niti preveč težavna), časovna omejitev oz. preostali čas pa morata biti med reševanjem izpisana na zaslonu, da množičnika spodbujata. Pomembno je tudi, da naloga vsebuje elemente naključnosti, npr. da naključno združuje množičnike v pare, naključno izbira besede, ki jih morajo prepoznati ipd. S tem poskrbimo, da naloga ni predvidljiva, obenem pa se izognemo morebitnemu goljufanju množičnikov.

5.2 Primeri mikronalog

V tem razdelku predstavljamo nekaj primerov različnih načinov množičenja (tako klasičnih mikronalog kot iger z namenom), ki so se izkazali kot uspešni v sorodnih raziskavah.

5.2.1 Reševanje mikronalog

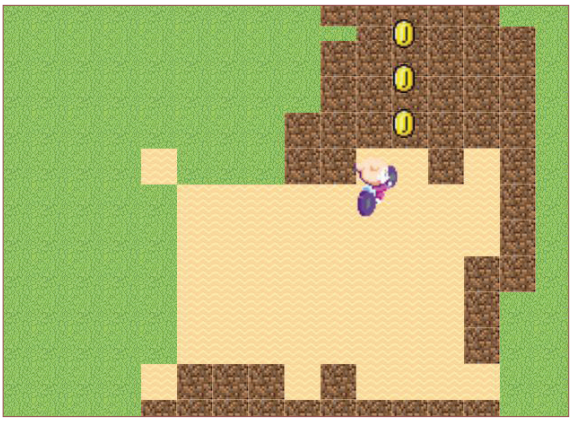
Slika 1 prikazuje primer mikronaloge za označevanje semantičnih vlog (Fossati et al. 2013). Naloga sestoji iz kratkega navodila, ki mu sledi poved, v kateri mora množičnik označiti vršilca dejanja (angl. *agent*) in del telesa (angl. *body part*). V tem primeru sta pravilna odgovora on (angl. *he*) in ni (angl. *none*).



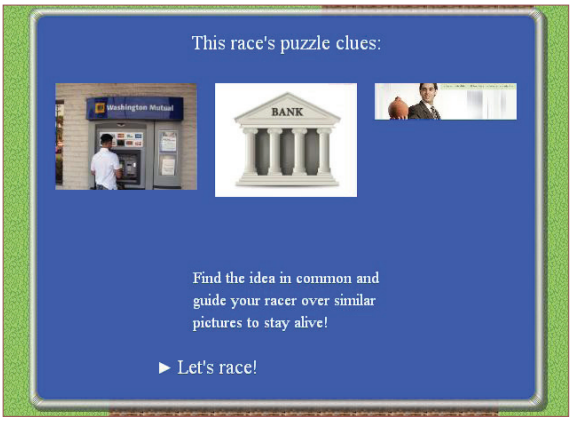
Slika 1: Mikronaloga za označevanje semantičnih vlog.

5.2.2 Reševanje mikronalog v igrah z namenom

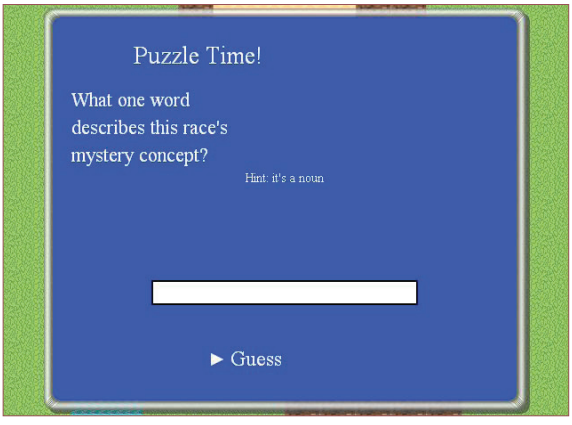
Slika 2 prikazuje vmesnik igre z namenom Puzzle Racer (Jurgens in Navigli 2014), pri kateri igralec tekmuje z avtomobilčkom ter pobira kovance in druge dobrine, ki prinašajo točke. Pred začetkom dirke igralec dobi namig v obliki treh slik (Slika 3), na podlagi katerih mora ugotoviti, kaj imajo skupnega, da lahko reši okvirček z uganko, ki se pojavi med dirkanjem (Slika 4). V tem primeru je pravilni odgovor denar (angl. *money*).



Slika 2: Igra z namenom Puzzle Racer.



Slika 3: Namig pri igri Puzzle Racer.



Slika 4: Uganka pri igri Puzzle Racer.

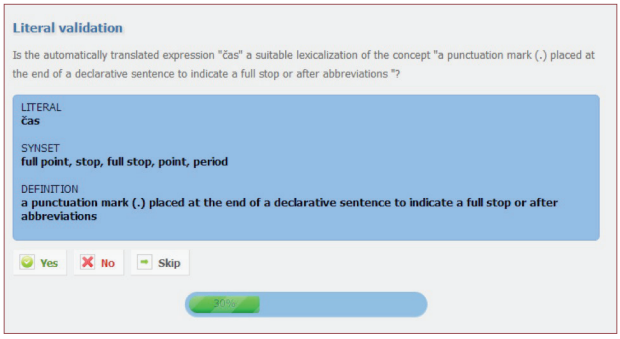
5.2.3 Reševanje mikronalog na družbenih omrežjih

Igre z namenom je mogoče vključiti tudi v družbena omrežja, kjer so prikladno dostopne velikemu številu uporabnikov, ki na njih preživljajo precej prostega časa in so dovzetni za tovrstne izzive. Na Sliki 5 je posnetek igre Phrase Detectives (Chamberlain et al. 2008) na družbenem omrežju Facebook. Igralec dobi zgled z dvema obarvanima besednima zvezama, od katerih se ena nanaša na drugo, označiti pa mora, ali se z oznakama strinja. Za pravilne odgovore (glede na ujemanje z drugimi igralci) prejme točke. V tem primeru je pravilen odgovor da (angl. *Agree*).

Slika 5: Različica igre Phrase Detectives na omrežju Facebook.

5.2.4 Primeri mikronalog za slovenščino

Na sliki 6 je prikaz mikronaloge v primeru množičenja za čiščenje sloWNeta (Fišer et al. 2014) z orodjem sloWCrowd (Tavčar et al. 2012). Množičnik mora pri nalogi označiti, ali dani literal (beseda ali besedna zveza) glede na angleške ustreznice in definicijo sodi v ta sinset (množico sinonimov). V tem primeru je pravilen odgovor ne (angl. *No*).



Slika 6: Mikronaloga za potrjevanje literalov v orodju sloWCrowd.

Slika 7 prikazuje vmesnik Igre besed. Igralec dobi iztočnico (v tem primeru pridevnik *gasilska*), za katero mora v 30 sekundah predlagati tri tipične kolokatorje. V ugibanju se lahko pomeri tudi z izbranim ali naključnim nasprotnikom. Odgovori se točkujejo glede na ujemanje s seznami kolokacij iz korpusa Gigafida.

Slika 7: Vmesnik Igre besed.

6 PREVERJANJE IN ZAGOTAVLJANJE KAKOVOSTI

V tem razdelku predstavljamo mehanizme, s katerimi lahko posredno ali neposredno zagotovimo kakovostne rezultate pri množičenju ter izločimo šumne prvine, ki se v množici rezultatov znajdejo npr. zaradi nejasnih navodil ali zaradi nezanesljivih množičnikov.

6.1 Zlati standard

Najpogostejša metoda nadziranja kakovosti se izvaja s pomočjo zlatega standarda (angl. *gold standard*), ročno označene množice podatkov, ki jo pri razvoju jezikovnih tehnologij uporabljamo za učno ali testno množico. Pri množičenju zlati standard vsebuje določeno število mikronalog, ki jih je vnaprej pravilno rešil strokovnjak. Naloge iz zlatega standarda so nato naključno vključene med mikronaloge, ki jih rešujejo množičniki, in služijo preverjanju njihove zanesljivosti – če množičnik ne odgovori pravilno na dovolj mikronalog iz zlatega standarda, se vsi njegovi odgovori izključijo iz končnih rezultatov.

Pri oblikovanju zlatega standarda je treba zagotoviti njegovo reprezentativnost, tako po obsegu kot težavnosti. S preveč preprostim zlatim standardom namreč

ne moremo učinkovito ločiti zanesljivih množičnikov od nezanesljivih, preveč težaven zlati standard pa bo izključil preveliko število množičnikov. Prav tako je treba pri reševanju mikronalog zagotoviti ustrezno razmerje vprašanj iz zlatega standarda in pravih vprašanj, saj s premajhnim številom vprašanj iz zlatega standarda ne moremo zanesljivo preveriti množičnikove zanesljivosti, s prevelikim številom vprašanj iz zlatega standarda pa namesto dragocenih odgovorov na nova vprašanja zbiramo že znane odgovore, kar je neekonomično.

6.2 Ujemanje med množičniki

Druga metoda za nadziranje kakovosti množičenja je izračun ujemanja med množičniki (angl. *inter-annotator agreement*). To storimo tako, da več različnim množičnikom ponudimo v reševanje iste mikronaloge. Na ta način pridobimo večje število odgovorov za vsako nalogo, odgovore pa lahko potem primerjamo in na podlagi primerjave odgovorov različnih množičnikov na isto vprašanje ugotovimo, kolikšno je razhajanje med njimi. Glede na porazdelitev odgovorov lahko izračunamo tudi stopnjo zanesljivosti (angl. *confidence score*) za posamezno mikronalogo ali množičnika (Oyama et al. 2013).

Če je veliko primerov, v katerih se odgovori množičnikov močno razhajajo, lahko to pomeni, da mikronaloge niso bile ustrezno zasnovane, da jih ni reševala skupina s pravo mero predznanja ali da smernice za označevanje niso bile dovolj jasno opredeljene, zaradi česar jih je treba izboljšati. V primerih, ko razhajanje ni pretirano, se lahko upošteva večinski glas (angl. *majority vote*), ki končno odločitev sprejme tako, da upošteva rešitev večine množičnikov.

Poudariti je treba, da je pomembno najti optimalno zgornjo mejo večkratnega postavljanja istega vprašanja različnim množičnikom, saj z vsakim ponovljenim vprašanjem ne dobimo nobenega novega odgovora, kar nezanemarljivo poveča stroške množičenja. Običajno so za eno odločitev potrebne 3 oznake, za bolj zapletene naloge pa 5 (tako predvidevajo tudi Krek et al. 2013b).

6.3 Razsojanje

Razsojanje (angl. *refereeing*) je proces, v katerem lahko težavne primere, pri katerih množičniki niso prišli do enotne rešitve, označi strokovnjak – razsodnik. Če je bila priprava na množičenje (z oblikovanjem mikronalog in smernic za označevanje) uspešna, je na tak način strokovnjakom na koncu prepuščen le manjši delež težavnih primerov, večina dela pa je še vedno množičena. V primeru označevanja

hrvaškega korpusa (Klubička in Ljubešić 2014) se je izkazalo, da tovrstni postopek količino dela strokovnjakov skoraj razpolovi.

6.4 Doslednost množičnika

Doslednost (angl. *consistency* ali *intra-annotator agreement*) je zadnja od priljubljenih metod, s katero iz rezultatov izločimo nezanesljive množičnike. Pri merjenju doslednosti namreč množičniku v reševanje ponudimo večkrat isto nalogo (Gut in Bayerl 2004), s čimer lahko preverimo, ali so njegovi odgovori konsistentni. Če se odgovori na iste naloge preveč razhajajo, množičnikovih rezultatov ne upoštevamo, saj bodisi ni dovolj samozavesten oziroma usposobljen bodisi izbira naključne odgovore, da bi rešil čimveč nalog.

7 SOOČANJE S PREDSODKI

Kljub veliki pozornosti, ki jo zadnje čase dobiva množičenje tudi med leksikografi, v razpravah o uporabi množičenja v leksikografiji še vedno naletimo na številne predsodke.[17] V tem razdelku naslavljamo poglavitne pomisleke in skušamo odpraviti najpogostejše dvome, ki so ovira pri uvedbi množičenja v leksikografske projekte.

7.1 Slovarja ne morejo pisati nestrokovnjaki

Danes je pri gradnji slovarskih priročnikov področje, na katerem so uporabniki slovarjev najbolj udeleženi, kolaborativna leksikografija, pri kateri uporabniki aktivno prispevajo iztočnice, razlage ipd. Ker so tudi nekateri avtorji z opredelitvijo množičenja nekoliko nedosledni (prim. Estellés-Arolas in González-Ladrón-de-Guevara 2012), se množičenje pogosto neupravičeno zamenjuje ali celo enači s kolaborativno leksikografijo.

Za razliko od številnih kolaborativnih projektov, pri katerih vse delo opravijo nestrokovnjaki, pri množičenju vedno aktivno sodeluje tudi pobudnik množičenja (izvajalec projekta), in sicer s pripravo podatkov, oblikovanjem mikronalog, preverjanjem kakovosti, zagotavljanjem motivacije med množičniki ipd. Prav tako drugače kot pri številnih kolaborativnih projektih, ki sicer dokazujejo, da lahko tudi uporabniki prispevajo h koristnim in široko uporabljanim slovarskim izdelkom (Meyer in Gurevych 2012), množičenje, kot ga predlagamo za vključitev v izdelavo slovarja sodobnega slovenskega jezika, zajema predvsem čiščenje

17 http://www.sssj.si/pogosta-vprasanja/ (dostop 8. 8. 2015).

avtomatsko luščenih podatkov pred gradnjo slovarja in ne predvideva kolaborativnega pristopa, pri katerem je takoj objavljen vsak uporabniški prispevek, množica pa neposredno nadzira tudi nabor besed, vključenih v slovar, vsebino in organizacijo informacij v geselskem članku (členitev pomenov, vrstni red definicij ipd.), kar je za končni izdelek lahko problematično. Meyer in Gurevych (2012) sicer ugotavljata, da kolaborativni projekti predstavljajo vsoto mnenj številnih avtorjev, ki geselske članke intenzivno popravljajo, vse dokler ni dosežen splošni konsenz tako o njihovi strukturi kot vsebini, zaradi česar kolaborativna leksikografija v številnih segmentih daje popolnoma primerljive rezultate resnim leksikografskim projektom. Kot največjo pomanjkljivost izpostavita učinkovit mehanizem za ločevanje med zrelimi, kakovostno oblikovanimi geselskimi članki in tistimi, ki še potrebujejo izboljšave. Podobno opozarja Lew (2013), ki opaža, da je pri določenih geslih vrstni red definicij v Wikislovarju precej naključen, pri čemer so lahko pri vrhu tudi povsem marginalni pomeni. Podobno je pri slovarju Urban Dictionary, pri katerem uporabniki z glasovanjem vplivajo na vrstni red definicij, sporno dejstvo, da lahko uporabniki izglasujejo tudi definicijo, ki se jim zdi najbolj zabavna oziroma ki najbolje odraža njihovo ideološko prepričanje, ni pa nujno najustreznejša.

Nasprotno je množičenje le ena od faz izdelave slovarja: najprej jezikovni tehnologi avtomatsko izluščijo podatke iz korpusa in drugih podatkovnih zbirk, ki jih množičniki s pomočjo ciljnih mikronalog očistijo, nakar jih leksikografi uporabijo pri ročnem leksikografskem delu. Množičenje je torej vmesni člen med avtomatsko in ekspertno obdelavo podatkov, saj z avtomatskim luščenjem in z delom množičnikov bistveno razbremeni leksikografa, hkrati pa vključuje ročni pristop pri postopkih, ki jih še ni mogoče zadovoljivo avtomatizirati. Ta metoda v obsežnem leksikografskem okolju še ni bila temeljito preizkušena, a na podlagi številnih drugih projektov, v katerih se je kot postopek za čiščenje avtomatsko generiranih podatkov izkazala kot učinkovita (Klubička in Ljubešić 2014; Fišer et al. 2014; Kosem et al. 2013), sklepamo, da bo uspešna tudi v leksikografiji.

7.2 Množičen slovar je nezanesljiv

Pogost predsodek zadeva tudi zanesljivost rezultatov množičenja, največkrat zato, ker so v delo (lahko) vključeni nestrokovnjaki. Na tem mestu je treba poudariti, da je serija mikronalog izdelana za določen profil množičnika glede na predznanje, ki je potrebno za reševanje naloge. Pri kakovostno zasnovanem postopku množičenja bodo bolj zapletene naloge reševali množičniki z več znanja (npr. študenti ali diplomanti jezikoslovnih smeri), preproste naloge pa bodo prepuščene tudi nestrokovnjakom.

V prispevku smo predstavili tudi vrsto mehanizmov, s katerimi lahko preverjamo in nadziramo kakovost pridobljenih rezulatov (npr. zlati standard, ujemanje med množičniki, večinski glas, doslednost, razsojanje) ter učinkovito izločimo tiste množičnike, ki dajejo nepravilne ali nezanesljive odgovore. Številni avtorji (prim. Rumshisky 2011; Fišer et al. 2014; Klubička in Ljubešić 2014; Fossati et al. 2013) so te mehanizme že preizkusili in ugotovili, da zagotavljajo visoko natančnost množičenjskih rezultatov, ki so enako kakovostni, kot če bi delo opravljali samo strokovnjaki (Snow et al. 2008).

7.3 Množičenje degradira leksikografski poklic

Množičenje kot nova oblika (prekarnega) dela, ki še ni izrecno predvidena v zakonodaji niti v tujini niti v Sloveniji, vzbuja tudi številne etične pomisleke, ki zadevajo predvsem plačilo množičnikov, pogoje dela in priznanje avtorstva. Pogosto uporabljane platforme za množičenje so največkrat le posrednice pobudnikov množičenja, ki ceno za opravljanje nalog na svojem projektu določijo sami. Množičnikom sicer ni treba sprejemati slabo plačanih nalog, a so v to pogosto prisiljeni, če želijo priti do zaslužka. Kot smo že omenili, na nizka plačila in izkoriščevalsko ravnanje z množičniki opozarjajo številni avtorji (Sabou et al. 2014; Silberman et al. 2010; Lease in Alonso 2014; Felstiner 2011), na tovrstne prakse nizkih plačil pa naletimo tudi pri sorodnih raziskavah: Snow et al. (2008) na platformi *Amazon Mechanical Turk* namreč plačajo skupno le 2 dolarja za 7.000 oznak nestrokovnjakov oziroma 1 dolar za 1.500 oznak strokovnjakov.

Dolžnost koordinatorjev vsakega leksikografskega projekta je torej, da množičnike kot vse ostale delavce obravnavajo korektno in jim zagotovijo ustrezno plačilo, kar je treba upoštevati že pri sami zasnovi projekta, ko se določa proračun. Obenem je treba poskrbeti, da je prispevek množičnikov na končnem izdelku tudi ustrezno priznan.

Poleg samega plačila in pogojev dela se pri množičenju pogosto soočamo tudi z mnenjem, da z izkoriščevalsko obliko prekarnega dela degradira poklic leksikografov in jezikoslovcev ter jim celo jemlje delo, ki ga preusmerja na (slabo plačano) nekvalificirano množico. Poudariti je treba, da je bistvo množičenja smiselno izkoriščanje virov – da se izurjenim leksikografom prihrani delo in dragoceni čas pri rutinskih opravilih, množičnikom pa omogoči, da po svojih močeh prispevajo h gradnji jezikovnih virov in v zameno dobijo motivacijo v različnih oblikah (denarno ali materialno plačilo, pridobivanje izkušenj in referenc, zabava ipd.).

7.4 Množičenje je sanjska rešitev

Za zagotavljanje ustrezne vloge množičenja v leksikografskih projektih je nujno prepoznati potencial, pa tudi omejitve množičenja, saj množičenje ni uporabno za vsako vrsto podatkov, vsako fazo leksikografskega dela in vsak leksikografski projekt. Množičenja recimo ne moremo uporabiti, če ne moremo zagotoviti rednega upravljanja s kampanjo (priprave mikronalog, preverjanja zbranih odgovorov, sprotne motivacije in plačevanja množičnikov). V vsebinskem smislu množičenje prav tako ni primerno za vprašanja odprtega tipa in vprašanja, ki zahtevajo podajanje subjektivnih ocen. Koristno je lahko le takrat, ko v leksikografskem projektu kljub vsem potrebnim pripravljalnim, vmesnim in naknadnim opravilom prihrani čas in/ali denar, pri tem pa še vedno zagotavlja zanesljive rezultate.

7.5 Ukrepi za zmanjšanje tveganj

Ker omenjeni predsodki pred množičenjskimi kampanjami niso prisotni samo v stroki, temveč nanje naletimo tudi pri splošni javnosti, je zelo pomembno, da ima pobudnik množičenja izdelano strategijo odnosov z javnostmi ter do potencialnih množičnikov pristopi pazljivo in premišljeno, hkrati pa od množice pričakuje vložek, ki je sorazmeren z vrsto predvidene motivacije. V primeru, ko množica za svoje delo ni plačana, ji na primer ni primerno zastavljati preveč ambicioznih nalog. Pomembno je tudi, da pobudnik množičenja skozi celotno množičenjsko kampanjo ohranja stik s skupnostjo množičnikov, jo obvešča o poteku projekta, vabi na javne predstavitve projekta in podobne dogodke, se jim javno zahvali za doprinos k projektu ipd.

8 ZAKLJUČEK

Številni jezikoslovni projekti so množičenje že uspešno uporabili na različnih področjih, kar kaže, da bi bila ta metoda lahko uporabna tudi v slovenski leksikografiji kot učinkovit način za obdelavo podatkov za gradnjo slovarja. Pri tem je treba vse potrebne vidike upoštevati že pri zasnovi leksikografskega projekta: od priprave podatkov, oblikovanja mikronalog in rekrutiranja množičnikov do zagotavljanja njihove motivacije in upoštevanja pravnih, finančnih ter etičnih omejitev projekta.

Da je množičenje mogoče uspešno uporabiti tudi v slovenskem okolju, dokazujejo dosedanje izkušnje pri sorodnih raziskavah, obenem pa je treba poudariti, da je tudi motiviranost slovenske javnosti za tovrstne projekte visoka. To kaže

npr. slovar pogovorne slovenščine Razvezani jezik, nezanemarljiv pa je tudi porast skupin z jezikoslovno tematiko na družbenih omrežjih, v katerih uporabniki zelo aktivno in redno sodelujejo.

Množičenje bo v okviru naslednje generacije leksikografskih projektov nedvomno postalo koristno orodje leksikografov, saj bo pripomoglo k hitrejšemu delu v obdobju, ko zahteve po hitri obdelavi vse večje količine jezikovnih podatkov naraščajo, in razbremenilo leksikografe pri rutinskih opravilih, zaradi česar jim bo ostalo več časa in energije za strokovno delo. Podatkovne baze, ki bodo rezultat množičenja, bodo imele dodano vrednost, saj jih bo mogoče uporabiti tudi za druge, neslovarske namene, npr. za izboljšanje jezikovnotehnoloških orodij z metodami strojnega učenja, pri čemer množičeni podatki služijo kot kakovostna učna množica. Slovar sodobnega slovenskega jezika je eden prvih leksikografskih projektov, ki ima v načrtu množičenje vključiti v celoten delotok, s čimer bo kot pionirski projekt začrtal smernice za številne prihodnje slovarje in jezikovne vire pri nas in po svetu.

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.

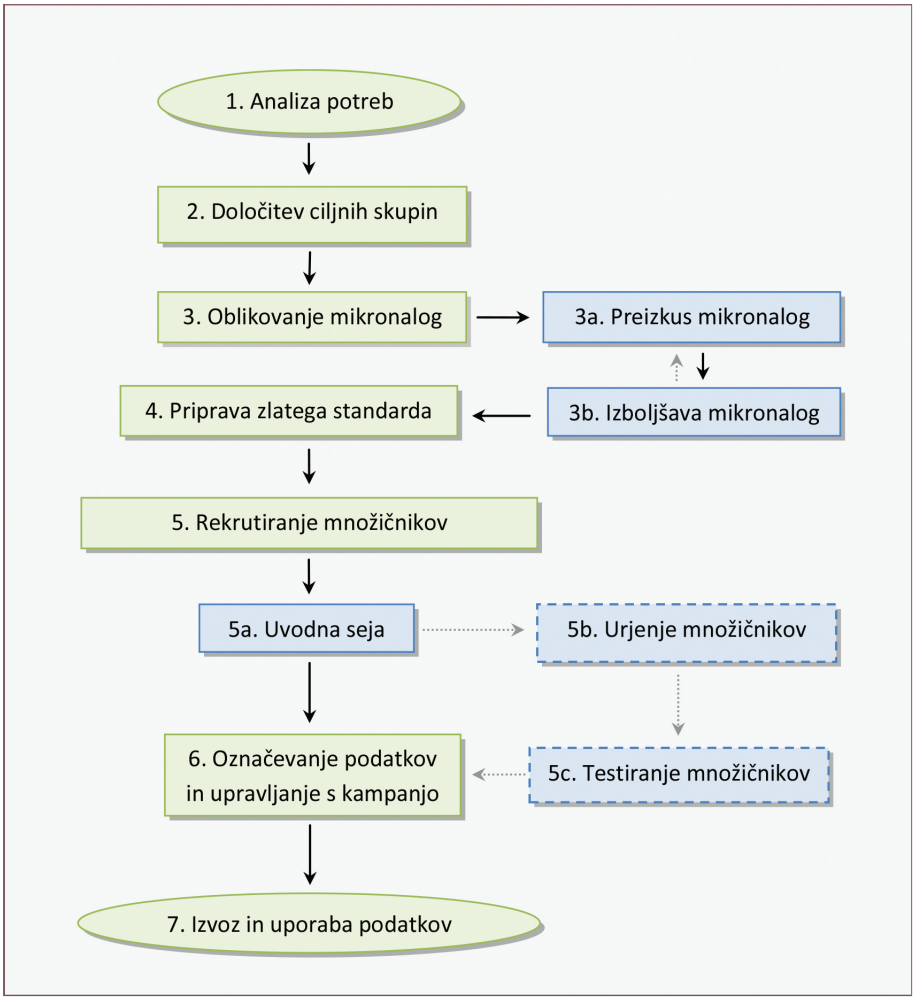
2 DELOTOK MNOŽIČENJA V LEKSIKOGRAFSKIH PROJEKTIH

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.

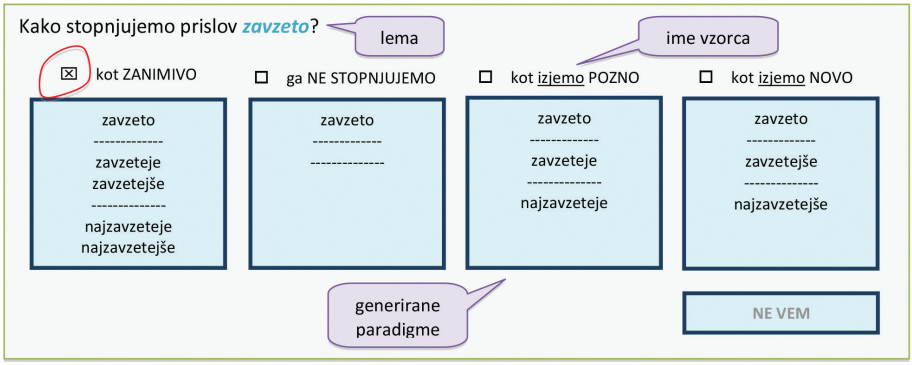
Za razliko od druge, jezikovnotehnološke vloge leksikona besednih oblik, ki v ospredje postavlja predvsem čim večjo pokritost besedišča, sta pri prvi, jezikovnopriročniški vlogi leksikona pomembna predvsem čim večja natančnost in zanesljivost prikazanih podatkov. To zahteva precejšnjo količino ročnega 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).

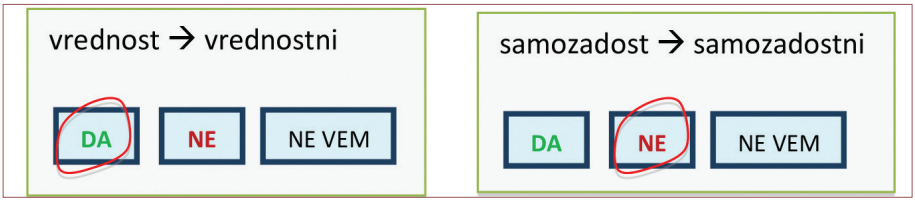
8 Pri tem med procesom avtomatskega luščenja (tj. iskanjem preseka generiranih paradigem in korpusnih podatkov) ter procesom množičenja dopuščamo možnost dodatnega avtomatskega filtriranja na obvladljivo število možnosti (npr. do največ tri paradigme), npr. s statističnimi izračuni verjetnosti glede 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).

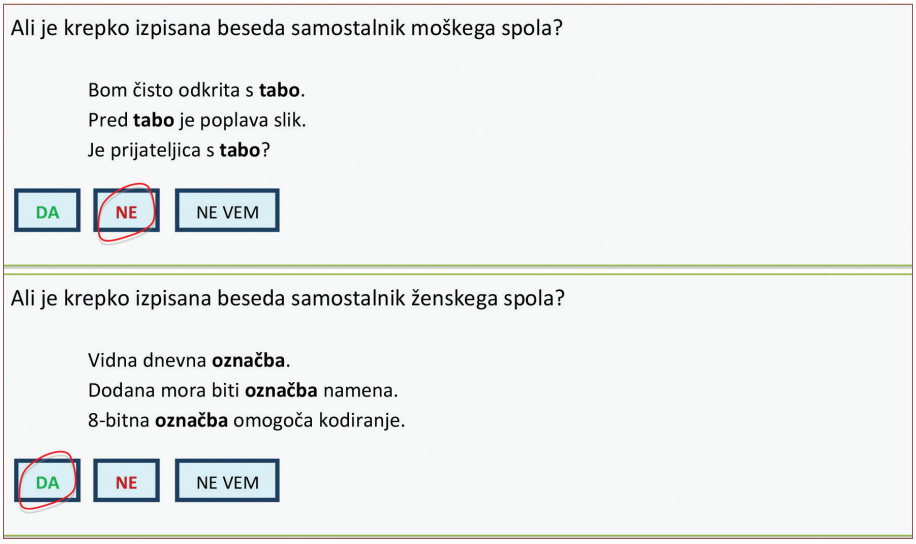


Slika 4: Primer uporabe množičenja za validacijo besedotvornih povezav.

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.

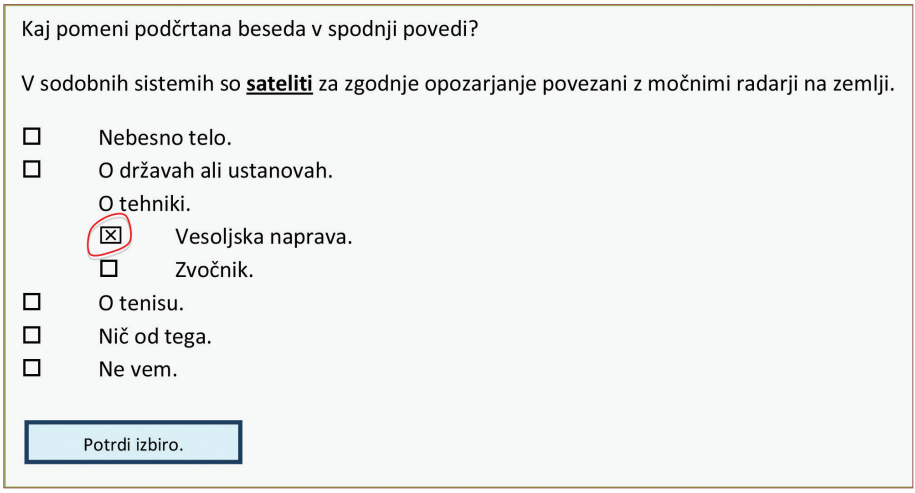
Dobra pomenska členitev je po našem mnenju taka, ki odraža čim večji konsenz v jezikovni skupnosti, zato predvidevamo, da bo prav analiza odgovorov, ki jih bodo prispevali množičniki nestrokovnjaki, omogočila prepoznavanje pomenskih opisov, ki so neproblematični in izkazujejo najvišjo stopnjo strinjanja, 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

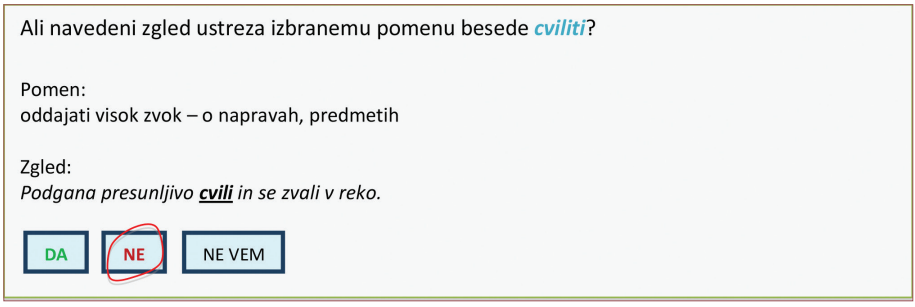
V prvi nalogi množičnikom ponudimo različne pomenske opise večpomenske besede, kot so zabeleženi v LBS, in jih prosimo, da zgled, v katerem je beseda

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.

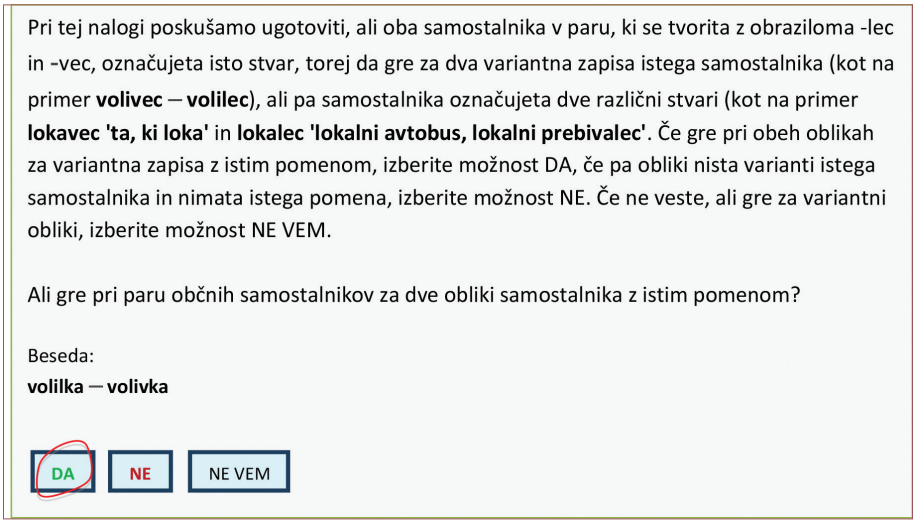
Oblikovanje mikronalog in zlatega standarda: V tem sklopu smo zaenkrat predvideli dve kampanji: določanje normativno zaznamovanih oblik in pregibanje tujejezičnih lastnih imen. Podatki za mikronaloge bodo na podlagi ročno določenih hevristik avtomatsko izluščeni iz korpusa, naloga množičnikov pa bo, 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.

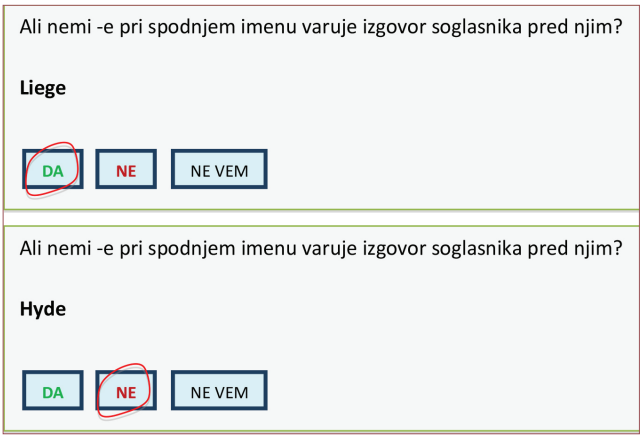


Slika 8: Iskanje parov variantnih lem znotraj kategorije D2c1a: Besedotvorje > Tvorba samostalnikov > Izbira priponskega obrazila > *-lec/-vec*.

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.

Literatura in viri

LITERATURA

Abel, Andrea in Christian M. Meyer, 2013: The dynamics outside the paper: user contributions to online dictionaries. Kosem, Iztok, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets in Maria Tuulik (ur.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference.* Ljubljana: Trojina, Institute for Applied Slovene Studies in Tallinn: Eesti Keele Instituut. 179–194.

Adam, Robert, 2015: *Morfologie.* Praha: Univerzita Karlova v Praze, Nakladatelství Karolinum.

Adamska-Sałaciak, Arleta, 2008: Prepositions in Dictionaries for Foreign Learners: A Cognitive Linguistic Look. Bernal, Elisenda in Janet DeCesaris (ur.): *Proceedings of the XIII Euralex International Congress. Barcelona: Universitat Pompeu Fabra.* 1477–1485.

Adolphs, Svenja in Ronald Carter, 2003: And she's like it's terrible, like: Spoken discourse, grammar and corpus analysis. *International Journal of English Studies* 3/1. 45–66.

Adolphs, Svenja in Ronald Carter, 2013: *Spoken Corpus Linguistics: From Monomodal to Multimodal.* New York in London: Routledge.

Ahačič, Kozma, 2007: *Zgodovina misli o jeziku in književnosti na Slovenskem: protestantizem*. Ljubljana: Založba ZRC, ZRC SAZU.

Ahlin, Martin, Branka Lazar, Zvonka Praznik in Jerica Snoj, 2014: Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja. *Jezik in slovstvo* 59/4. 121–127.

Ahmad, Khurshid, Willy Martin, Martin Hölter in Margaret Rogers, 1995: Specialist terms in general language dictionaries. *University of Surrey Technical Report CS-95-14*. http://www. mcs. surrey. ac. uk (dostop 14. 7. 2015).

Ahn, Luis von in Laura Dabbish, 2004: Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems*. New York: ACM. 319–326.

Ahn, Luis von, 2006: Games with a Purpose. *Computer* 39/6. 92–94.

Ahn, Luis von, Mihir Kedia in Manuel Blum, 2006a: Verbosity: a game for collecting common-sense facts. Grinter, Rebecca, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries in Gary Olson (ur.): *Proceedings of the SIGCHI conference on Human Factors in computing systems*. New York: ACM. 75–78.

Ahn, Luis von, Ruoran Liu in Manuel Blum, 2006b: Peekaboom: A Game for Locating Objects in Images. Grinter, Rebecca, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries in Gary Olson (ur.): *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM. 55–64.

Ahn, Luis von in Laura Dabbish, 2008: Designing Games with a Purpose. *Communications of the ACM* 51/8. ACM. 58–67.

Ahn, Luis von, 2013. Duolingo: learn a language for free while helping to translate the web. *Proceedings of the international conference on Intelligent user interfaces*. New York: ACM. 1–2.

Aijmer, Karin, 1996: *English Discourse Particles: Evidence from a Corpus*. Amsterdam in Philadelphia: John Benjamins.

Akhmanova, Olga, 1976: *Linguostylistics. Theory and Method*. Berlin in New York: de Gruyter.

Akinnaso, F. Niyi, 1982: On the differences between spoken and written language. *Language and Speech* 25/2. 97–125.

Akkaya, Cem, Alexander Conrad, Janyce Wiebe in Rada Mihalcea, 2010: Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation. *Proceedings of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*. Los Angeles, California, ZDA. Association for Computational Linguistics. 195–203.

Al-Ajmi, Hashan, 2008: The Effectiveness of Dictionary Examples in Decoding: The Case of Kuwaiti Learners of English. *Lexikos* 18. 15–26.

Anward, Jan, 2001: Parts of speech. Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher in Wolfgang Raible (ur.): *Language Typology and Language Universals* 20/1. Berlin in New York: de Gruyter. 726–735.

Apresjan, Juri, D., 1973: Regular Polysemy. *Linguistics* 142. 5–39.

Apresjan, Juri, D., 2002: Principles of Systematic Lexicography. Corréard, Marie-Helene (ur.): *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. UK: Euralex. 91–104.

Arhar Holdt, Špela in Vojko Gorjanc, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

Arhar, Špela, 2009: Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54/3-4. 43–56.

Arhar, Špela in Peter Holozan, 2009: Leksikalna podatkovna zbirka ASES (Amebisov skupni elektronski slovar). Mikolič, Vesna (ur.): *Jezikovni korpusi v medkulturni komunikaciji.* Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Založba Annales in Zgodovinsko društvo za južno primorsko. 30−51.

Arhar Holdt, Špela, Gaja Červ, Polona Gantar, Iztok Kosem, Karmen Kosem, Irena Krapš Vodopivec, Simon Krek, Sara Može, Tadeja Rozman, Ana Marija Sobočan, Mojca Stritar Kučuk in Ana Zwitter Vitez, 2013a: *Pedagoški slovnični portal.* Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://slovnica.slovenščina.eu/ (dostop 12. 6. 2015).

Arhar Holdt, Špela, Kaja Dobrovoljc in Damjan Popič, 2013b: Reprezentacija standardnega in nestandardnega v virih SSJ. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 19–27.

Arhar Holdt, Špela, 2015: Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi koraki. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 136–149.

Arhar Holdt, Špela in Tadeja Rozman, 2015: *Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov*. V pripravi.

Arhar Holdt, Špela, Jaka Čibej in Ana Zwitter Vitez, 2015: S pomočjo uporabniških jezikovnih vprašanj in mnenj do boljšega slovarja. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 126–214.

Atkins, B. T. Sue in Antonio Zampolli (ur.), 1994: *Computational approaches to the lexicon.* Oxford: Oxford University Press.

Atkins, B. T. Sue, 1996: Bilingual Dictionaries. Past, Present and Future. Gellerstam, Martin, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström in Catarina Röjder Papmehl (ur): *EURALEX '96 Proceedings*. Göteborg: Univerza v Göteborgu. 515–546.

Atkins, B. T. Sue (ur.), 1998: *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators.* Tübingen: Max Niemeyer Verlag.

Atkins, B. T. Sue in Krista Varantola, 1998: Monitoring dictionary use. Atkins, B. T. Sue (ur.): *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Tübingen: Max Niemeyer Verlag. 83–122.

Atkins, B. T. Sue in Michael Rundell, 2008: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Atkins, B. T. Sue, Adam Kilgarriff in Michael Rundell, 2010: Database of Analysed Texts of English (DANTE): the NEID database project. Dykstra, Anne in Tanneke Schoonheim (ur.): *Proceedings of the XIV Euralex International Conference*. Leeuwarden, Netherlands: Fryske Akademy. 549–556.

Aust, Ronald, Mary Jane Kelley in Warren Roby, 1993: The Use of Hyper-Reference and Conventional Dictionaries. *Educational Technology Research and Development* 41/4. 63–73.

Bajec, Anton, Rudolf Kolarič in Mirko Rupel, 1956 (1971): *Slovenska slovnica.* Ljubljana: DZS.

Balažic Bulc, Tatjana, 2009: Slovarski opisi pomensko občutljivih kategorij: konektorji in SSKJ. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 33–39

Balažic Bulc, Tatjana, 2015: Členek v slovenskem jezikoslovju in slovarju. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 524–538.

Barnbrook, Geoff, 2002: *Defining language: A local grammar of definition sentences*. Amsterdam in Philadelphia: John Benjamins.

Barnhart, Clarence L., 1962: Problems in editing commercial monolingual dictionaries. *International Journal of American Linguistics* 28/2. 161–181.

Baroni, Marco in Silvia Bernardini, 2004: BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of the 4th Language Resources and Evaluation Conference (LREC).* Lizbona. 1313–1316.

Baroni, Marco, Adam Kilgarriff in Jan Pomikálek, 2006: WebBootCaT: Instant Domain-Specific Corpora to Support Human Translators. *Proceedings of EAMT*. Oslo. 247–252.

Battenburg, John, 1989: *A Study of English Monolingual Learners' Dictionaries and their Users*. Ph.D. Dissertation. Purdue University, IN, ZDA.

Behrens, Bergljot, 2008: Explaining advanced L2: Discourse structural properties of coordinating conjunction in English L1 and advanced L2. Ramm, Wiebke in Cathrine Fabricius-Hansen (ur.): *Linearisation and Segmentation in Discourse. Multidisciplinary Approaches to Discourse 2008.* Oslo: Department of Literature, Area Studies and European Languages, University of Oslo. 17–29.

Béjoint, Henri, 1981: The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. *Applied Linguistics* II/3. 207–222.

Béjoint, Henri, 1988: Scientific and technical words in general dictionaries. *International journal of lexicography* 1/4. 354–368.

Béjoint, Henri, 2000: *Modern Lexicography. An Introduction*. Oxford: Oxford University Press.

Béjoint, Henri, 2007: Nouvelle lexicographie et nouvelles termininologies: convergences et divergences. L'Homme, Claude-Marie in Sylvie Vandaele: *Lexicographie et terminographie*. Ottawa: University of Ottawa Press. 29–78.

Bentivogli, Luisa, Marcello Federico, Giovanni Moretti in Michael Paul, 2011: Getting expert quality from the crowd for machine translation evaluation. *Proceedings of the 13th Machine Translation Summit*. Xiamen, Kitajska. 521–528.

Bergenholtz, Henning in Sven Tarp, 2000: The concept of dictionary usage. *Nordic Journal of English Studies* 3.1. 23–36.

Bergenholtz, Henning in Sven Tarp, 2003: Two Opposing Theories: On H. E. Wiegand's Recent Discovery of Lexicographic Functions. *Hermes. Journal of Linguistics* 31. 171–196.

Bergenholtz, Henning in Mia Johnsen, 2005: Log files as a tool for improving Internet dictionaries. *Hermes. Journal of Linguistics* 34. 117–141.

Bergenholtz, Henning, 2011: Access to and presentation of needs-adapted data in monofunctional internet dictionaries. Fuertes-Olivera, Pedro A. in Henning Bergenholtz (ur.): *E-lexicography: the Internet, digital initiatives and lexicography*. London in New York: Continuum. 30–45.

Bezlaj, France, 1939: *Oris slovenskega knjižnega jezika*. Ljubljana: Znanstveno društvo.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad in Edward Finegan, 1999: *Longman Grammar of Spoken and Written English*. Longman.

Biber, Douglas, 2009: A corpus-driven approach to formulaic language: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14/3. 381–417.

Biber, Douglas, 2012: Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8/1. 9–37.

Biemann, Chris in Valerie Nygaard, 2010: Crowdsourcing WordNet. *Proceedings* of the *5th Global WordNet Conference*. Mumbaj, Indija.

Bishop, Jonathan, 2009: Enhancing the Understanding of Genres of Web-Based Communities: the Role of the Ecological Cognition Framework. *Int. J. Web Based Communities* 5/1. 4–17.

Bizjak Končar, Aleksandra, Helena Dobrovoljc, Kaja Dobrovoljc, Nataša Logar Berginc, Polonca Kocjančič, Simon Krek in Tadeja Rozman, 2011: Slogovni priročnik – projekt *»Sporazumevanje v slovenskem jeziku« – Kazalnik 17*. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik17/Kazalnik_17_Slogovni_prirocnik_SSJ.pdf (dostop 30. 6. 2015).

Blei, David M., Andrew Y. Ng in Michael I. Jordan, 2003: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3. 993–1022.

Bogaards, Paul, 1998: What Type of Words do Language Learners Look Up? Atkins, B. T. Sue (ur.): *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag. 151–157.

Bogaards, Paul, 2003: Uses and users of dictionaries. Van Sterkenburg, Piet (ur.): *A Practical Guide to Lexicography.* Amsterdam in Philadelphia: John Benjamins. 26–33.

Boguraev, Bran in Ted Briscoe, 1987: Large lexicons for natural language processing: Utilising the grammar coding system of LDOCE. *Computational Linguistics* 13/3-4. 203–18.

Boguraev, Bran in Ted Briscoe (ur.), 1989: *Computational Lexicography for Natural Language Processing*. London in New York: Longman.

Bolinger, Dwight, 1965: The atomization of meaning. *Language* 41/4. 555–573.

Boonmoh, Atipat, 2012: E-dictionary Use under the Spotlight. Students' Use of Pocket Electronic Dictionaries for Writing. *Lexikos* 22. 43–68.

Boulanger, Jean-Claude in Marie-Claude L'Homme, 1991: Les technolectes dans la pratique dictionnairique générale. Quelques fragments d'une culture. *Meta: Journal des traducteurs* 36/1. 23–40.

Boulanger, Jean-Claude, 1996: Les Dictionnaires généraux monolingues, une voie royale pour les technolectes. *TradTerm* 3. 137–151.

Bourdieu, Pierre, 1982: *Ce que parler veut dire*. Pariz: Fayard.

Brants, Thorsten, 2000: TnT: a statistical part-of-speech tagger. *Proceedings of the sixth conference on Applied natural language processing*. Seattle, Washington: Association for Computational Linguistics. 224–231.

Burzio, Luigi, 1994: *Principles of English stress*. Cambridge: Cambridge University Press.

Buzássyová, Klára, 2009: Slovar sodobnega slovaškega jezika (Z vidika zasnove in organizacije dela). Perdih, Andrej (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 119–124.

Caluwe, Johan de in Johan Taeldeman, 2003: Morphology in dictionaries. Sterkenburg, Piet van (ur.)*: A practical guide to lexicography*. Amsterdam in Philadelphia: John Benjamins. 114–126.

Cazinkić, Robert, 2012: Opombe k oblikoslovnemu delu Toporišičeve Slovenske slovnice (2000). *Jezikoslovni zapiski* 18/1. 179–205.

Chafe, Wallace in Jane Danielewich, 1987: *Properties of spoken and written language. Technical report no. 5.* Berkeley: University of California, Center for the Study of Writing.

Chamberlain, Jon, Massimo Poesio in Udo Kruschwitz, 2008: Phrase Detectives: A Web-based collaborative annotation game. Ghidini, Chiara, Axel-Cyrille Ngonga Ngomo, Stefanie Lindstaedt in Tassilo Pellegrini (ur.): *Proceedings of the 7th International Conference on Semantic* Systems, I-*SEMANTICS*. New York: ACM.

Channell, Joanna, 1999: Corpus-based Analysis of Evaluative Texts. Hunston, Susan in Geoff Thompson (ur.): *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press. 39–55.

Chen, Yuzhen, 2010. Dictionary use and EFL learning. A contrastive study of pocket electronic dictionaries and paper dictionaries. *International Journal of Lexicography* 23/3. 275–306.

Christ, Oli, 1994: A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94 proceedings*. Budapest. 23–32.

Cook, Guy, 1992 (2006): *The Discourse of Advertising.* London in New York: Routledge.

Cook, Paul, Michael Rundell, Jey Han Lau in Timothy Baldwin, 2014: Applying a word-sense induction system to the automatic extraction of dictionary examples'. Abel, Andrea, Chiara Vettori, Natascia Ralli (ur.): *Proceedings of the XVI EURALEX International Congress.* Bolzano, Italy: EURAC. 319–328.

Cooper, Robert L., 1989: *Language Planning and Social Change*. Cambridge: Cambridge University Press.

Corris, Miriam, Christopher Manning, Susan Poetsch in Jane Simpson, 2000: Bilingual Dictionaries for Australian Languages: User studies on the place of paper and electronic dictionaries. Heid, Ulrich, Stefan Evert, Egbert Lehmann in Christian Rohrer (ur.): *Proceedings of the Ninth EURALEX International Congress, Stuttgart, Germany, August 8th-12th 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung. 169–181.

Cosme, Christelle, 2006: Clause combining across languages: A corpus-based study of English-French translation shifts. *Languages in Contrast*. 6/1. 71–108.

Crowston, Kevin, 2010: Internet Genres. *Encyclopedia of Library and Information Sciences*. New York: CRC Press. http://crowston.syr.edu/sites/crowston.syr.edu/files/elischapter.pdf (dostop 24. 6. 2015).

Crystal, David, 1987 (1997): *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.

Crystal, David, 1995: *The Cambridge Encyclopedia of the English Language.* Cambridge: Cambridge University Press.

Crystal, David, 2001 (2006): *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, David, 2011: *Internet Linguistics: A Student Guide*. Oxon in New York: Routledge.

Cvrček, Václav, Vilém Kodýtek, Marie Kopřivová, Dominika Kováříková, Petr Sgall, Michal Šulc, Jan Táborský, Jan Volín in Martina Waclawičová, 2010: *Mluvnice současné češtiny* 1. Praha: Univerzita Karlova v Praze, Nakladatelství Karolinum.

Čebulj, Monika: *Raba slovarja v 1. in 2. triletju osnovne šole*. Diplomsko delo. Ljubljana: Pedagoška fakulteta UL.

Čechová, Marie in kolektiv, 1997: *Stylistika současné češtiny.* Praha: ISV – nakladatelství.

Čermák, František, 1995: Jazikový korpus: Prostredek a zdroj poznání. *Slovo a slovesnost* 56/2. 119–140.

Čermák, František, 2001: *Jazyk a jazykověda.* Praha: Univerzita Karlova v Praze, Nakladatelství Karolinum.

Černelič, Ivana, 1991: Členek kot besedna vrsta v slovenskem knjižnem jeziku. *Jezikoslovni zapiski* 1. 73–85.

Černelič-Kozlevčar, Ivanka, 1988: Reševanje besednovrstnih vprašanj v Slovarju slovenskega knjižnega jezika. Paternu, Boris in Franc Jakopin (ur.): *Sodobni slovenski jezik, književnost in kultura. Obdobja* 8. Ljubljana: Filozofska fakulteta, Znanstveni inštitut, Oddelek za slovanske jezike in književnosti. 289–300.

Černelič-Kozlevčar, Ivanka, 1993: O delitvi členkov. Orožen, Martina in Mateja Hočevar (ur.): *Vprašanja slovarja in zdomske književnosti. Zborovanje slavistov v Murski Soboti*. Ljubljana: Zavod RS za šolstvo. 213–225.

Čibej, Jaka, Darja Fišer in Iztok Kosem, 2015a: The role of crowdsourcing in lexicography. Kosem, Iztok, Miloš Jakubiček, Jelena Kallas in Simon Krek (ur.): *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of eLex 2015, 11–13 August 2015, Herstmonceux Castle, UK*. Ljubljana in Brighton: Trojina, Institute for Applied Slovene Studies in Lexical Computing Ltd. 70–83.

Čibej, Jaka, Vojko Gorjanc in Damjan Popič, 2015b: Vloga jezikovnih vprašanj prevajalcev pri načrtovanju novega enojezičnega slovarja. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 168–181.

Čmejrková, Svetla, 2000: *Reklama v češtine.* Praha: Leda.

Dardano, Maurizio in Pietro Trifone, 1995: *Grammatica italiana con nozioni di linguistica.* Bologna: Zanichelli.

Davies, Alan, 1997: Real Language Norms: description, prescription and their critics. A Case for Applied Linguistics. *VIEWS: Vienna English Working Papers*. 4–18. http://www.univie.ac.at/Anglistik/views/views972.pdf (dostop 17. 7. 2015).

De Cock, Sylvie, 2002: Pragmatic prefabs in learners‘ dictionaries. Braasch, Anna in Claus Povlsen (ur.): *Proceedings of the Tenth EURALEX International Congress. EURALEX 2002*. København: Center for Sprogteknologi. 471–481.

De Schryver, Gilles-Maurice, 2003: Lexicographers‘ Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16/2. 143–199.

De Schryver, Gilles-Maurice, David Joffe, Pitta Joffe in Sarah Hillewaert, 2006: Do dictionary users really look up frequent words?—on the overestimation of the value of corpus-based lexicography. *Lexikos* 16/1. 67–83.

Declaration on Acces to Research Data from Public Funding. http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157 (dostop 6. 7. 2015).

Denkowski, Michael in Alon Lavie, 2010: Exploring normalization techniques for human judgments of machine translation adequacy collected using Amazon Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10*. Stroudsburg, ZDA: Association for Computational Linguistics. 57–61.

Dictionnaire du français contemporain, 1966. Pariz: Larousse.

Dictionnaire du français d'aujourd'hui, 2000. Pariz: Larousse.

Dictionnaire du français usuel, 2007. Bruxelles: De Boeck.

Dictionnaire Maxi Débutants, 1997. Pariz: Larousse.

Dijk, Teun A. van, 1979: Pragmatic connectives. *Journal of Pragmatics* 3/5. 447–456.

Dijk, Teun A. van, 1993: Principles of critical discourse analysis. *Discourse & Society* 4/2. 249–283.

Dijk, Teun A. van, 2001: Critical discourse analysis. Tannen, Deborah, Deborah Schiffrin in Heidi E. Hamilton (ur.): *Handbook of discourse analysis*. Oxford: Blackwell. 352–371.

Dilts, Philip in John Newman, 2006: A note on quantifying 'good' and 'bad' prosodies. *Corpus Linguistics and Linguistic Theory* 2. 233–242.

Dobrovoljc, Helena, 2004: *Pravopisje na Slovenskem*. Ljubljana: Založba ZRC, ZRC SAZU.

Dobrovoljc, Helena in Nataša Jakop, 2011: *Sodobni pravopisni priročnik med normo in predpisom*. Ljubljana: Založba ZRC, ZRC SAZU.

Dobrovoljc, Helena in Simon Krek, 2011: Normativne zadrege – empirični pristop. Kranjc, Simona (ur.): *Meddisciplinarnost v slovenistiki. Obdobja* 30. Ljubljana: Znanstvena založba Filozofske fakultete UL. 89–97.

Dobrovoljc, Helena, 2014: Normativna informacija v slovarju. Grahek, Irena in Simona Bergoč (ur.): *Novi slovar za 21. Stoletje. Posvet o novem slovarju slovenskega jezika, Ministrstvo za kulturo, 12. februar 2014*. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/19-Helena_Dobrovoljc-prispevek-posvet-oddano.pdf (dostop 4. 8. 2015).

Dobrovoljc, Kaja, Simon Krek in Jan Rupnik, 2012: Skladenjski razčlenjevalnik za slovenščino. *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 42–47.

Dobrovoljc, Kaja in Simon Krek, 2013: Spletni portal Slogovni priročnik: luščenje in prikaz podatkov o jezikovni rabi. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 101–107.

Dobrovoljc, Kaja, Simon Krek, Peter Holozan, Tomaž Erjavec in Miro Romih, 2013: Morphological lexicon Sloleks 1.2. *Slovenian language resource repository CLARIN.SI.* http://hdl.handle.net/11356/1039 (dostop 30. 6. 2015).

Dobrovoljc, Kaja, 2014: Re-evaluating morphological dictionaries: the case of adverbs in Slovene. *International NooJ 2014 Conference*. Sassari, Italija.

Dobrovoljc, Kaja, 2015: Oblikoslovne informacije v sodobnih slovarskih priročnikih. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 64–79.

Dobrovoljc, Kaja, Simon Krek in Tomaž Erjavec, 2015: Leksikon besednih oblik Sloleks in smernice njegovega razvoja. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 80–105.

Dolar, Kaja, 2014: Kolaborativni slovar Razvezani jezik. *Slavistična revija* 62/2. 235–252.

Domingo, David in Ari Heinonen, 2008: Weblogs and Journalism: a Typology to Explore the Blurring Boundaries. *Nordicom Review* 29/1. 3–15.

Dubois Jean in Claude Dubois, 1971: *Introduction à la lexicographie: le dictionnaire.* Pariz: Larousse.

Dubois, Claude, 1990: Considérations générales sur l'organisation du travail lexicographique. Hausmann, Franz J., Oskar Reichmann, Herbert E. Wiegand in Ladislav Zgusta (ur.): *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie.* Berlin in New York: de Gruyter. 1574–1588.

Dürscheid, Christa, 2007: *Syntax. Grundlagen und Theorien.* Göttingen: Vandenhoeck & Ruprecht.

Dziemianko, Anna, 2010: Paper or electronic? The role of dictionary form in language reception, production and the retention of meaning and collocations. *International Journal of Lexicography* 23/3. 257–273.

Dziemianko, Anna, 2015: Colours in Online Dictionaries: A Case of Functional Labels. *International Journal of Lexicography* 28/1. 27–61.

El-Haj, Mahmoud, Udo Kruschwitz in Chris Fox, 2014: Creating Language Resources for Under-resourced Languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation*. Springer.

Epple, Barbara, 2000: Sexismus in Wörterbüchern. Heid, Ulrich, Stefan Evert, Egbert Lehmann in Christian Rohrer (ur.): *Proceedings of the 9th EURALEX International Congress*. Stuttgart: Universität Stuttgart. 739–749.

Erjavec, Tomaž, Nancy Ide, Vladimir Petkević in Jean Véronis, 1995: Multilingual Text Tools and Corpora for Central and Eastern European Languages. *Language resources for language technology: proceedings of the first European seminar.* Tihany, Madžarska.

Erjavec, Tomaž, 1998: Oznake korpusa FIDA. Štrukelj, Inka (ur.): *Jezik za danes in jutri*. Ljubljana: Društvo za uporabno jezikoslovje Slovenije in Inštitut za narodnostna vprašanja. 85–95.

Erjavec, Tomaž, Vojko Gorjanc in Marko Stabej, 1998: Korpus FIDA. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 124–127.

Erjavec, Tomaž in Sašo Džeroski, 2004: Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words. *Applied artificial intelligence* 18. 17–41.

Erjavec, Tomaž, Matija Ogrin in Jože Faganel, 2004: E-Slomšek: a TEI encoding of a critical edition of 19th century Slovenian rhetoric prose. *New Technologies and standards: Digitization of national heritage*, junij 3-5, 2004, Beograd. Pregled Nacionalnog centra za digitalizaciju 5. 31–41.

Erjavec, Tomaž, Camelia Ignat, Bruno Pouliquen in Ralf Steinberger, 2005: Massive Multi Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences* 15. 529–540.

Erjavec, Tomaž in Simon Krek, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 46–53.

Erjavec, Tomaž, Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček, 2008: *Specifikacije za leksikon (besednih oblik) – projekt »Sporazumevanje v slovenskem jeziku« – kazalnik 3*. http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik3/SSJ_Kazalnik_3_Specifikacije-leksikon_v1.pdf (dostop 30. 6. 2015).

Erjavec, Tomaž, 2009: Odprtost jezikovnih virov za slovenščino. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 115–121.

Erjavec, Tomaž in Simon Krek, 2010: *Training corpus jos1M 1.1. Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1037 (dostop 15. 1. 2015).

Erjavec, Tomaž, Darja Fišer, Simon Krek in Nina Ledinek, 2010a: Jezikovni viri projekta JOS. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Sedme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 42–48.

Erjavec, Tomaž, Simon Krek, Špela Arhar, Darja Fišer, Nina Ledinek, Amanda Saksida, Breda Sivec in Blaž Trebar, 2010b: *Priporočila za jezikovno označevanje JOS*. Dodatek B.9 Členek. http://nl.ijs.si/jos/msd/html-sl/back.1_div.2_div.9.html (dostop 15. 6. 2015).

Erjavec, Tomaž in Nikola Ljubešić, 2011: hrWac in slWaC: Compiling Web Corpora for Croatian and Slovene. Habernal, Ivan in Václav Matoušek (ur.): *Text, Speech and Dialogue. Proceedings of the 14th International Conference, TSD*. Pilsen: Springer Berlin Heidelberg. 395–402.

Erjavec, Tomaž, 2012a: Jezikovni viri starejše slovenščine IMP: zbirka besedil, korpus, slovar. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 52–56.

Erjavec, Tomaž, 2012b: MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation* 46/1. 131–142.

Erjavec, Tomaž in Nataša Logar Berginc, 2012: Referenčni korpusi slovenskega jezika (cc) Gigafida in (cc)KRES. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 57–62.

Erjavec, Tomaž, 2013a: Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0* 1/1. 24–49. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf (dostop 6. 7. 2015).

Erjavec, Tomaž, 2013b: Slovene corpora for corpus linguistics and language technologies. Gajdošová, Katarína in Adriána Žáková (ur.): *Natural Language Processing. Corpus Linguistics. e-Learning*. Lüdenscheid: RAM-Verlag. 51–61.

Erjavec, Tomaž, 2014: Odprt dostop do podatkovne baze slovarja. Grahek, Irena in Simona Bergoč (ur.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo.* Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/20-_Tomaz_Erjavec--SlovarPosvet.pdf (dostop 6. 7. 2015).

Erjavec, Tomaž in Nikola Ljubešić, 2014: The slWaC 2.0 Corpus of the Slovene Web. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Devete konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 50–55.

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

Erjavec, Tomaž, Peter Holozan in Nikola Ljubešić, 2015a: Jezikovne tehnologije in zapis korpusa. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 262–276.

Erjavec, Tomaž, Darja Fišer, Nikola Ljubešić, Nataša Logar in Vesna Mikolič, 2015b: Nadgradnja Gigafide: spletna besedila. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 242–261.

Erjavec, Tomaž, Nikola Ljubešić in Nataša Logar, 2015c: The slWaC Corpus of the Slovene Web. *Informatica* 39/1. 35–42.

Erlandsen, Jens, 2004: iLex – new DWS. *Third International Workshop on Dictionary Writing systems (DWS 2004)*, *Brno, 6.–7. September 2004.*

Estellés-Arolas, Enrique in Fernando González-Ladrón-de-Guevara, 2012: Towards an Integrated Crowdsourcing Definition. *Journal of Information Science* 38/2. 189–200.

Fairclough, Norman, 1989 (2001): *Language and power.* London in New York: Longman.

Felstiner, Alek, 2011: Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley Journal of Employment and Labor Law*. 143–203.

Ferbežar, Ina, Mihaela Knez, Andreja Markovič, Nataša Pirih Svetina, Mojca Schlamberger Brezar, Marko Stabej, Hotimir Tivadar in Jana Zemljarič Miklavčič, 2004: *Sporazumevalni prag za slovenščino*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete UL in Ministrstvo RS za šolstvo, znanost in šport.

Fillmore, Charles J., Collin F. Baker in Hiroaki Sato, 2004: FrameNet as a „Net". *Proceedings of LREC* 4. Lisbon: ELRA. 1091–1094.

Finkel, Jenny Rose, Trond Grenager in Christopher Manning, 2005: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. 363–370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf (dostop 15. 1. 2015).

Finkel, Jenny Rose, Christopher Manning in Andrew Y. Ng, 2006: Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, *Sydney, Australia, July 22-23*. 618–626.

Fishman, Joshua A., 1968: Language problems and types of political and sociocultural integration: A conceptual postscript. Fishman, Joshua A., Charles A. Ferguson in Jyotirindra das Gupta (ur.): *Language problems of the developing nations*. New York, London, Sydney in Toronto: John Wiley & Sons, Inc. 491–498.

Fishman, Joshua A., 1989: *Language and ethnicity in minority sociolinguistic perspective.* Clevedon in Philadelphia: Multilingual Matters Ltd.

Fišer, Darja, 2009: sloWNET – slovenski semantični leksikon. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 145–149.

Fišer, Darja, Aleš Tavčar in Tomaž Erjavec, 2014a: sloWCrowd: A crowdsourcing tool for lexicographic tasks. *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14*. 4371–4375.

Fišer, Darja, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešić, 2014b: JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Devete konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 56–61.

Fišer, Darja in Jaka Čibej, 2015: Potencial množičenja v sodobni leksikografiji. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 565–586.

Fišer, Darja, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Iztok Kosem, Špela Arhar Holdt, Damjan Popič in Tomaž Erjavec, 2015: Množičenje za slovar sodobnega slovenskega jezika. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. XXX–XXX.

Flowerdew, Lynne, 2011: *Corpora and Language Education*. Palgrave.

Fort, Karen, Gilles Adda, Benoît Sagot, Joseph Mariani in Alain Couillault, 2014: Crowdsourcing for Language Resource Development: Criticisms About Amazon Mechanical Turk Overpowering Use. *Human Language Technology Challenges for Computer Science and Linguistics*. Springer. 303–314.

Fossati, Marco, Claudio Giuliano in Sara Tonelli, 2013: Outsourcing FrameNet to the Crowd. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofija, Bolgarija. Association for Computational Linguistics. 742–747.

Fox, Gwyneth, 1987: The Case for Examples. Sinclair, John McH.(ur.): *Looking up: An Account of the COBUILD Project in Lexical Computing*. London: Collins. 137–149.

Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet in Claudia Soria, 2006: Lexical Markup Framework (LMF). *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06*. Pariz: ELRA. 233–236.

Frankenberg-Garcia, Ana, 2012: Learners' Use of Corpus Examples. *International Journal of Lexicography* 25/3. 273–296.

Frankenberg-Garcia, Ana, 2014: The Use of Corpus Examples for Language Comprehension and Production. *ReCALL* 26. 128–146.

Frath, Pierre, 2000: Polysemy, Homonymy and Reference. Albert Hamm (ur.): *Proceedings of the JASGIL Seminar, Strasbourg 5-6 May 2000*. Strasbourg: RANAM, Recherches Anglaises et Nord-Américaines. 43–56.

Fuertes-Olivera, Pedro A. in Sven Tarp, 2014: *Theory and practice of specialised online dictionaries: Lexicography versus terminography*. Berlin in New York: de Gruyter.

Galisson, Robert, 2001: Une dictionnairique à géométrie variable au service de la lexiculture. Pruvost, Jean: *Les dictionnaires de langue fran*çaise. Pariz: Honoré Champion. 115–138.

Gantar, Polona, 2007: *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: Založba ZRC, ZRC SAZU.

Gantar, Polona, 2009: Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo* 54/3-4. 69–94.

Gantar, Polona in Simon Krek, 2009: Drugačen pogled na slovarske definicije: opisati, pojasniti, razložiti? Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete. 151–159.

Gantar, Polona, Katja Grabnar, Polonca Kocjančič, Simon Krek, Olga Pobirk, Rok Rejc, Mojca Šorli, Simon Šuster in Petra Zaranšek, 2009: Specifikacije za izdelavo leksikalne baze za slovenščino – projekt *»Sporazumevanje v slovenskem jeziku« – Kazalnika 5 in 6*. http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik5/SSJ_Kazalnik_5_Specifikacije-opis-analize-korpusa_v1.pdf in http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik6/SSJ_Kazalnik_6_Specifikacije-leksikalna-baza_v1.pdf (dostop 19. 9. 2015).

Gantar, Polona, 2010: K uporabniku usmerjeni slovnično-leksikalni opisi slovenskega jezika. Gorjanc, Vojko in Andreja Žele (ur.): *Izzivi sodobnega jezikoslovja*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 35–51.

Gantar, Polona, 2011: Leksikalna baza za slovenščino: komu, zakaj in kako (naprej)? *Jezikoslovni zapiski* 17/2. 77–92.

Gantar, Polona in Simon Krek, 2011: Slovene lexical database. Majchráková, Daniela in Radovan Garabík (ur.): *Natural language processing, multilinguality*. Brno: Tribun EU. 72–80.

Gantar, Polona, Iztok Kosem, Simon Krek in Mojca Šorli, 2011: *Leksikalna baza za slovenščino. Navodila za avtorje*, julij 2011. Projekt *»Sporazumevanje v slovenskem jeziku«*. Kamnik. Interno gradivo.

Gantar, Polona, Simon Krek, Iztok Kosem, Mojca Šorli, Katja Grabnar, Olga Pobirk, Petra Zaranšek in Nina Drstvenšek, 2012: *Leksikalna baza za slovenščino*. Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. http://www.slovenscina.eu/spletni-slovar/leksikalna-baza, http://www.slovenscina.eu/spletni-slovar/prenos (dostop 9. 8. 2015).

Gantar, Polona in Iztok Kosem, 2013: Beleženje in prikazovanje podatkov o jezikovni rabi: od leksikalne baze do spletnega slovarja. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 133–139.

Gantar, Polona, 2014: Slovar sodobnega slovenskega jezika: leksikografska tradicija in/ali inovacija. *Slovenščina 2.0* 2/2. 194–231. http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_10.pdf (dostop 8. 8. 2015).

Gantar, Polona, 2015a: Homonimija in večpomenskost: od teorije do slovarja, 2015: Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. XXX–XXX.

Gantar, Polona, 2015b: *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete UL. V pripravi.

Gantar, Polona, Iztok Kosem in Simon Krek, 2015a: Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 280–297.

Gantar, Polona, Iztok Kosem, Simon Krek in Vojko Gorjanc, 2015b: Collocations dictionary of Slovene: challenge for automatization and crowdsourcing. Corpas Pastor, Gloria, Miriam Buendía Castro in Rut Gutierrez Florido (ur.): *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives.* Europhras 2015, Malaga, 29 June to 1 July 2015.

Gao, Qin in Stephan Vogel, 2010: Consensus versus expertise: a case study of word alignment with Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10*. Stroudsburg, ZDA: Association for Computational Linguistics. 30–34.

Gliha Komac, Nataša, Nataša Jakop, Janoš Ježovnik, Simona Klemenčič, Domen Krvina, Nina Ledinek, Tanja Mirtič, Andrej Perdih, Špela Petric, Marko Snoj in Andreja Žele, 2015: *Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika*. Različica 1.1. Ljubljana: Inštitut za slovenski jezik Frana Ramovša; Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti, 2015. http://www.fran.si/179/novi-slovar-slovenskega-knjiznega-jezika/datoteke/Koncept_NoviSSKJ.pdf (dostop 30. 6. 2015).

Goddard, Angela in Lindsey Meân Patterson, 2000: *Language and Gender*. London in New York: Routledge.

Golden, Marija, 2000: *Teorija opisnega jezikoslovja* 1. *Skladnja.* Ljubljana: Filozofska fakulteta.

Golik, Pavel, Zoltán Tüske, Ralf Schlüter in Hermann Ney, 2013: Development of the RWTH Transcription System for Slovenian. Bimbot, Frédéric (ur.): *14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013): Speech in Life Sciences and Human Societies. Proceedings*. International Speech Communication Association (ISCA). 3107–3111.

Gorjanc, Vojko, 1998: Konektorji v slovničnem opisu znanstvenega besedila. *Slavistična revija* 46/4. 367–388.

Gorjanc, Vojko, 1999: Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. Križšnik, Erika (ur.): *35. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete. 47–59.

Gorjanc, Vojko, 2000: Nekatere možnosti jezikoslovne izrabe enojezikovnih korpusov. Orel, Irena (ur.): *36. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete. 335–348.

Gorjanc, Vojko, 2004: Politična korektnost in slovarski opisi slovenščine – zgolj modna muha? Stabej, Marko (ur.): *Moderno v slovenskem jeziku, literaturi in kulturi. 40. seminar slovenskega jezika, literature in kulture.* Ljubljana: Filozofska fakulteta. 153–161.

Gorjanc, Vojko, 2005a: Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. *Družboslovne razprave* 21/48. 197–209.

Gorjanc, Vojko, 2005b: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Gorjanc, Vojko, Simon Krek in Polona Gantar, 2005: Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50/2. 3–19.

Gorjanc, Vojko, 2009: Jezikovnotehnološka podpora slovarskemu delu. Perdih, Andrej (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 45–52.

Gorjanc, Vojko, 2012: Ideologija heteronormativnosti, prevodna in jezikovna norma. Bjelčevič, Aleš (ur.): *Ideologije v slovenskem jeziku, literaturi in kulturi. 40. seminar slovenskega jezika, literature in kulture*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. 38–44.

Gorjanc, Vojko in Darja Fišer, 2013: *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske fakultete.

Gorjanc, Vojko, 2014a: O heteronormativnosti slovarskega opisa slovenskega jezika: homoseksualnost, ekshibicionizem in druge perverznosti. *Narobe* 7/27-28. 12–15.

Gorjanc, Vojko, 2014b: Slovar slovenskega jezika v digitalni dobi. Grahek, Irena in Simona Bergoč (ur.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo*. Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/1-_Vojko_Gorjanc_-_Slovar_MK_tekst_FINAL.pdf (dostop: 15. 6. 2015).

Gorjanc, Vojko, Simon Krek in Damjan Popič, 2015: Med ideologijama knjižnega in standardnega jezika. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 32–48.

Górski, Rafał L. in Marek Łazinski, 2012: Typologia tekstów w NKJP. Przepiórkowski, Adam et al. (ur.): *Narodowy korpus języka polskiego.* Varšava: Wydawnictwo Naukowe PWN. 13–23.

Grabnar, Katarina in Mojca Šorli, 2008: Sodobno slovensko dvojezično slovaropisje (ob Velikem angleško-slovenskem slovarju Oxford-DZS in Malem slovensko-angleškem slovarju DZS). Jesenšek, Marko (ur.): *Od Megiserja do elektronske izdaje Pleteršnikovega slovarja*. Maribor: Filozofska fakulteta, Oddelek za slovanske jezike in književnosti. 339–360.

Gramley, Stephan in Kurt-Michael Pätzold, 1992: *A Survey of Modern English Grammar.* London: Routledge

Granda, Stane, 1999: *Prva odločitev Slovencev za Slovenijo*. Ljubljana: Založba Nova revija.

Grčar, Miha, Simon Krek in Kaja Dobrovoljc, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 89–94.

Greenbaum, Sidney, 1988: *Good English and the grammarian*. London in New York: Longman.

Grošelj, Robert, 2015: Besedne vrste v slovenskem jeziku. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 498–513.

Gut, Ulrike in Petra Saskia Bayerl, 2004: Measuring the Reliability of Manual Annotations of Speech Corpora. *The Proceedings of Speech Prosody 2004*. Nara, Japan. 565–568.

Haase, Peter, Jeen Broekstra, Andreas Eberhart in Raphael Volz, 2004: A Comparison of RDF Query Languages. McIlraith, Sheila A., Dimitris Plexousakis in Frank van Harmelen (ur.): *The Semantic Web – ISWC 2004*. Berlin in Heidelberg: Springer. 502–517.

Hajnšek-Holz, Milena, 1993. Leksikografski problemi prenosa knjižne oblike Slovarja slovenskega knjižnega jezika v računalniško. Štrukelj, Inka (ur.): *Jezik tako in drugače*. Ljubljana: Društvo za uporabno jezikoslovje Slovenije. 420–432.

Halle, Morris in Jean-Roger Vergnaud, 1987: *An essay on stress*. Cambridge, MA: MIT Press.

Halliday, M. A. K. in Ruqaiya Hasan, 1976: *Cohesion in English.* London in New York: Longman.

Halliday, M. A. K., 1978: *Language As Social Semiotic (The Social Interpretation of Language and Meaning)*. London: Edward Arnold.

Halliday, M. A. K., 1985: *An Introduction to Functional Grammar*. London: Edward Arnold.

Hanks, Patrick, 2004: Corpus Pattern Analysis. Williams, Geoffrey in Sandra Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004*. Lorient: Universite de Bretagne-sud. 87–97.

Hanks, Patrick in James Pustejovsky, 2005: A Pattern Dictionary for Natural Language Processing. *Revue Française de Linguistique Appliquée* 10/2. 63–82.

Hanks, Patrick, 2013: *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Hartmann, Reinhard R. K, 1987: Four perspectives on dictionary use: a critical review of research methods. Cowie, Anthony P. (ur.): *The Dictionary and the Language Learner*. Tübingen: Niemeyer. 11–28.

Hartmann, Reinhard R. K. in Gregory James, 1998: *Dictionary of Lexicography*. London: Routledge. E-knjiga. Hartmann, Reinhard R. K. (ur.): *Dictionaries in Language Learning: Recommendations, National Reports and Thematic Reports from the TNP Sub-Project 9: Dictionaries*. Berlin: Freie Universität. 36–52.

Hartmann, Reinhard R. K., 1999: The Exeter University Survey of Dictionary Use. Hartmann, Reinhard R. K. (ur.): *Dictionaries in Language Learning: Recommendations, National Reports and Thematic Reports from the TNP Sub-Project 9: Dictionaries.* Berlin: Freie Universität. 36–52.

Harvey, Keith in Deborah Yuill, 1997: A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. *Applied Linguistics* 18/3. 253–278.

Hatherall, Glyn, 1984: Studying dictionary use: some findings and proposals. Hartmann, Reinhard R. K. (ur.): *LEX'eter ,83 Proceedings: Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983*. Tübingen: Niemeyer Verlag. 183–189.

Haugen, Einar, 1968: The scandinavian languages as cultural artiface. Fishman, Joshua A., Charles A. Ferguson in Jyotirindra das Gupta (ur.): *Language problems of the developing nations*. New York, London, Sydney in Toronto: John Wiley & Sons, Inc. 267–284.

Haugen, Einar, 1983: The implementation of corpus planning: theory and practice. Cobarrubias, Juan in Joshua A. Fishman (ur.): *Progress in Language Planning: international perspectives*. Berlin: Mouton. 269–289.

Hayes, Bruce, 1995: *Metrical Stress Theory. Principles and case studies*. Chicago: University of Chicago Press.

Heinonen, Tarja, 2014: Workflow in Kielitoimiston sanakirja. *Workflow of Corpus-based Lexicography*, *COST ENeL WG3 meeting, Bolzano, 19 julij*. http://www.elexicography.eu/wp-content/uploads/2014/07/Heinonen_2014_COST_Bolzano.pdf (dostop 6. 7. 2015).

Henne, Helmut, 1972: *Semantik und Lexikographie. Untersuchungen zur lexikalischen Kodifikation der deutchen Sprache*. Berlin in New York: de Gruyter.

Heinrichsen, Peter Juel, Jens Allwood, 2005: Swedish and Danish, spoken and written language: A statistical comparison. *International Journal of Corpus Linguistics* 10/3. 367–399.

Herring, Sussan C., 2001: Computer-Mediated Discourse. Schiffrin, Deborah, Deborah Tannen in Heidi E. Hamilton (ur.): *The Handbook of Discourse Analysis.* Oxford: Blackwell Publishers. 612–634.

Herring, Susan C., Lois Ann Scheidt, Sabrina Bonus in Elijah Wright, 2004: Bridging the Gap: a Genre Analysis of Weblogs. *Proceedings of the 37th Hawaii International Conference on System Sciences*. IEEE – Institute of Electrical and Electronics Engineers. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.2930&rep=rep1&type=pdf (dostop 24. 6. 2015).

Herrity, Peter, 2000: *Slovene. A Comprehensive Grammar*. London in New York: Routledge.

Hinrichs, Erhard, Marie Hinrichs in Thomas Zastrow, 2010: WebLicht: Web-based LRT services for German. *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics. 25–29. http://www.aclweb.org/anthology/P10-4005 (dostop 15. 1. 2015).

Hirci, Nataša, 2003: Prevajanje danes in jutri: delo s sodobnimi prevajalskimi viri in orodji. *Jezik in slovstvo* 3-4/48. 89–102.

Hirci, Nataša, 2009: Empirične raziskovalne metode za opazovanje prevajalskega procesa. Pokorn, Nike K. (ur.): *Sodobne metode v prevodoslovnem raziskovanju*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 60–78.

Hirci, Nataša, 2013: Changing trends in the use of translation resources: the case of trainee translators in Slovenia. *ELOPE* 10. 149–165.

Hirci, Nataša in Tamara Mikolič Južnič, 2014: Korpusna raziskava rabe vzročnih in pojasnjevalnih povezovalcev v prevodih iz angleščine in italijanščine. Pisanski Peterlin, Agnes in Mojca Schlamberger Brezar (ur.): *Prevodoslovno usmerjene kontrastivne študije.* Ljubljana: Znanstvena založba Filozofske fakultete UL.

Hoekstra, Eric, 2010: Grammatical information in dictionaries. Dykstra, Anne in Tanneke Schoonheim (ur.): *Proceedings of the XIV EURALEX International Congress*. Afük, Ljouwert: Fryske Akademy. 1007–1012.

Hoey, Michael, 2005: *Lexical Priming*. London in New York: Routledge.

Hoffmanová, Jana, 1997: *Stylistika a ...* Praha: Trizonia.

Holozan, Peter, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček, 2008: *Specifikacije za učni korpus – projekt »Sporazumevanje v slovenskem jeziku« – Kazalnik 2*. http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik2/SSJ_Kazalnik_2_Specifikacije-ucni-korpus_v1.pdf (dostop 30. 6. 2015).

Holozan, Peter, 2011: *Samodejno izdelovanje besedilnih logičnih nalog v slovenščini.* Magistrsko delo. Ljubljana: Fakulteta za računalništvo in informatiko UL.

Holozan, Peter, 2012: Kako dobro programi popravljajo vejice v slovenščini. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 101−106.

Honselaar, Wim, 2003: Examples of design and production criteria for bilingual dictionaries. Sterkenburg, Piet van (ur.)*: A practical guide to lexicography*. Amsterdam in Philadelphia: John Benjamins. 323−332.

Horvat, Aleš in Jernej Vičič, 2012: Strojno prevajanje med slovenščino in španščino. Baldomir Zajc in Andrej Trost (ur.): *Zbornik enaindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2012*. Portorož, Slovenija. 101–104.

Householder, Fred W., 1967: Summary report. Householder, Fred W. in Sol Saporta (ur.): *Problems in lexicography*. Bloomington: Indiana University Publications. 279–282.

Howe, Jeff, 2008: *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. New York: Crown Publishing Group.

Hribar, Nataša, 2009: Al' prav se piše ... – spletna razglabljanja o slovenskem jeziku. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 171–176.

Hulst, Harry van der, 2010: Word accent: Terms, typologies and theories. Hulst, Harry van der, Rob Goedemans in Ellen van Zanten (ur.): *Stress patterns of the world. Part II: The data*. Berlin in New York: de Gruyter. 3–54.

Hult, Ann-Kristin, 2012: Old and New User Study Methods Combined–Linking Web Questionnaires with Log Files from the Swedish Lexin Dictionary. Vatvedt Fjeld, Ruth in Julie Matilde Torjusen (ur.): *Proceedings of the 15th EURALEX International Congress. EURALEX 2012*. Oslo: Universitetet i Oslo, Institutt for lingvistiske og nordiske studier. 922–928.

Hult, Ann-Kristin, 2014: The Authentic Voices of Dictionary Users – Viewing Comments on an Online Learner's Dictionary Before and After Revision. Abel, Andrea, Chiara Vettori, Natascia Ralli (ur.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano/Bozen. 237–247.

Humar, Marjeta, 2009: Terminologija v novem slovarju slovenskega jezika. Perdih, Andrej (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 61–74.

Humblé, Philippe, 2001: *Dictionaries and Language Learners*. Frankfurt am Main: Haag & Herchen.

Hunston, Susan in Sara Laviosa, 2001: *Corpus linguistics*. Birmingham: Centre for English Language Studies, The University of Birmingham.

Hunston, Susan, 2007: Semantic Prosody Revisited. *International Journal of Corpus Linguistics* 12/2. 249–268.

Ide, Nancy in Jean Véronis, 1994: MULTEXT: Multilingual Text Tools and Corpora. *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94*. Kyoto, Japan. 588–592.

ISO 24611, 2012: *Language resource management – Morpho-syntactic annotation framework (MAF)*.

Itô, Junko in Armin Mester, 2003: Weak layering and word binarity. Honma, Takeru, Masao Okazaki, Toshiyuki Tabata in Shin Ichi Tanaka (ur.): *A New Century of Phonology and Phonological Theory. A Festschrift for Professor Shosuke Haraguchi on the Occasion of his Sixtieth Birthday.* Tokyo: Kaitakusha. 26–65.

Itô, Junko in Armin Mester, 2009: The extended prosodic word. Grijzenhout, Janet in Barış Kabak (ur.): *Phonological domains: universals and deviations*. Berlin in New York: de Gruyter. 105–194.

Jackson, Howard, 1988: *Words and Their Meaning*. London: Longman.

Jackson, Howard, 2002: *Lexicography: an introduction*. Routeledge.

Jakop, Nataša, 2000: Vezanost poudarnih členkov na določeno besedno vrsto oziroma stavčni člen. *Jezikoslovni zapiski* 6. 67–80.

Jakop, Nataša, 2000/2001: Funkcijska delitev členkov: značilnosti naklonskih členkov. *Jezik in slovstvo* 46/7–8. 305–316.

Jakopin, Primož in Aleksandra Bizjak Končar, 1997: O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija* 45/3-4. 513-532.

Jakubíček, Miloš, Adam Kilgarriff, Vojtech Kovář, Pavel Rychlý in Vid Suchomel, 2013: The TenTen Corpus Family. *7th International Corpus Linguistics Conference CL2013*. Lancaster, UK. 125–127.

Javoršek, Jan Jona, 2015: *Razvoj korpusnega skladenjskega razčlenjevalnika.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Jeffries, Lessley, 2010: *Critical Stylistics: The Power of English*. Houndmills in New York: Palgrave Macmillan.

Jeffries, Lessley in Dan McIntyre, 2010: *Stylistics*. Cambridge: Cambridge University Press.

Jernudd, Björn in Jiři Nekvapil, 2012: History of the field: a sketch. Spolsky, Bernard (ur.): *The Cambridge Handbook of Language Policy.* Cambridge: Cambridge University Press. 16–36.

Jesenovec, Mojca, 2004: Poučevanje, učenje in pomnjenje leksike drugega/tujega jezika. *Jezik in slovstvo* 49/3-4. 35–47.

Jewler, A. Jarome in Bonnie L. Drewniany, 2005: *Creative strategy in Advertising*. Belmont: Thomson in Wadsworth.

Joseph, John E. in Talbot J. Taylor (ur.): *Ideologies of languages*. London: Routledge.

Josselin-Leray, Amelie, 2005: *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues.* Doktorska disertacija. Université Lumière Lyon II.

Josselin-Leray, Amelie in Roda P. Roberts, 2005: In Search of Terms: An Empirical Approach to Lexicography. *Meta: Journal des traducteurs* 50/4. 256–265.

Josselin-Leray, Amelie in Roberts, Roda P., 2007: La définition des termes dans les dictionnaires généraux unilingues: analyse de quelques exemples du domaine de la volcanologie à la lumière d'un corpus de vulgarisation. L'Homme, Marie-Claude in Sylvie Vandaele (ur.): *Lexicographie et terminologie: compatibilité des modèles et des méthodes*. Ottawa: Presses de l'Université d'Ottawa. 141–188.

Jošt, Kaja: *Večnaglasnice v slovenščini*. Magistrska naloga. Ljubljana: Filozofska fakulteta UL. V pripravi.

Joubert, Alain in Mathieu Lafourcade, 2012: A new dynamic approach for lexical networks evaluation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC'12*. Istanbul, Turky. 3687–3691.

Jurgec, Peter, 2000: Fonologija v novem slovenskem pravopisu. *Slava* 14. 45–63.

Jurgec, Peter, 2004: Fonologija v slovarju novejšega besedja. *Jezikoslovni zapiski* 10/2. 89–101.

Jurgec, Peter, 2005: Formant frequencies of Standard Slovenian vowels. *Govor* 22. 127–144.

Jurgec, Peter, 2006a: Formantne frekvence samoglasnikov v tonemski in netonemski standardni slovenščini. *Slavistična revija* 54/pos. št. 103–114.

Jurgec, Peter, 2006b: O nenaglašenih /e/ in /o/ v standardni slovenščini. *Slavistična revija* 54/2. 173–185.

Jurgec, Peter, 2007: *Novejše besedje s stališča fonologije: primer slovenščine*. Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Jurgec, Peter, 2010: O prihodnosti fonologije slovenščine in v Sloveniji. Gorjanc, Vojko in Andreja Žele (ur.): *Izzivi sodobnega jezikoslovja* Ljubljana: Znanstvena založba Filozofske fakultete UL. 13–34.

Jurgec, Peter, 2011: Slovenščina ima 9 samoglasnikov. *Slavistična revija* 59/3. 243–268.

Jurgec, Peter, 2015: Izgovor v slovarju sodobnega slovenskega jezika. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 382–391.

Jurgens, David in Roberto Navigli, 2014: It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics* 2. Association for Computational Linguistics. 449–463.

Juršič, Matjaž, Igor Mozetič, Tomaž Erjavec in Nada Lavrač, 2010: LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of universal computer science* 16/9. 1190–1214.

Kager, René, 2007: Feet and metrical stress. De Lacy, Paul (ur.): *The Cambridge handbook of phonology*. Cambridge: Cambridge University Press. 195–227.

Kalin Golob, Monika, 1996: *Jezikovna kultura in jezikovni kotički.* Ljubljana: Jutro.

Kalin Golob, Monika, 2013: Lačen si ful drugačen: v iskanju naslovnikovega jezika. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 201–206.

Kalin Golob, Monika, 2015: Stilistika v slovarju: oznaka publicistično. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis. Obdobja* 34. Ljubljana: Znanstvena založba Filozofske fakultete UL. V tisku.

Kallas, Jelena, Maria Tuulik in Margit Langemets, 2014: The Basic Estonian Dictionary: the First Monolingual L2 Learner's Dictionary of Estonian. *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano. http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_086_p_1109.pdf (dostop 6. 7. 2015).

Kallas, Jelena, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Margit Langemets, Jan Michelfeit, Maria Tuulik in Ülle Viks, 2015: Automatic generation of the Estonian Collocations Dictionary database. Kosem, Iztok, Miloš Jakubiček, Jelena Kallas in Simon Krek (ur.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of eLex 2015, 11–13 August 2015, Herstmonceux Castle, UK.* Ljubljana in Brighton: Trojina, Institute for Applied Slovene Studies in Lexical Computing Ltd. 1–20.

Karlík, Petr, Marek Nekula in Jana Pleskalová, 2002: *Enciklopedický slovník češtiny.* Praha: Nakladatelství Lidové noviny.

Katnić Bakaršić, Marina, 2007: *Stilistika*. Sarajevo: Ljiljan.

Katnić Bakaršić, Marina, 2012: *Između diskursa moći i moći diskursa*. Zagreb: Naklada ZORO.

Kay, Paul in Charles Fillmore, 1999: Grammatical constructions and linguistic generalizations: The What's X Doing Y construction? *Language* 75/1. 1–33.

Kennedy, Graeme, 1999: *An Introduction to Corpus Linguistics*. London in New York: Longman.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz in David Tugwell, 2004: The Sketch Engine. Williams, Geoffrey in Sandra Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004*. Lorient: Universite de Bretagne-sud. 105–116.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell in Pavel Rychly, 2008: GDEX: Automatically Finding Good Dictionary Examples in a Corpus. Bernal, Elisenda in Janet DeCesaris. (ur.): *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. 425–432.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý in Vít Suchomel, 2014: The Sketch Engine: ten years on. *Lexicography* 1/1. 7–36.

Kirkpatrick, Betty, 1985: A lexicographical dilemma: Monolingual dictionaries for the native: speaker and for the learner. Ilson, Robert (ur.): *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon Press. 7–13.

Klein, Wolfgang in Alexander Geyken, 2010: Das Digitale Wörterbuch der Deutschen Sprache (DWDS). Heid, Ulrich, Stefan Schierholz, Wolfgang Schweickard in Herbert Ernst Wiegand (ur.): *Lexikographica*. Berlin in New York: de Gruyter. 79–93.

Klemenc, Bojan, Marko Robnik-Šikonja, Luka Fürst, Ciril Bohak in Simon Krek, 2015: Tehnološka izvedba sodobnega digitalnega slovarja. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 52–63.

Klosa, Anette, 2013: The lexicographical process (with special focus on online dictionaries). Gouws, Rufus H., Ulrich Heid, Wolfgang Schweickard in Herberst Ernst Wiegand (ur.): *Dictionaries. An international Encyclopedia of Lexicography.* Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin in Boston: de Gruyter. 517–524.

Klubička, Filip in Nikola Ljubešić, 2014: Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of Croatian. *Jezikovne tehnologije. Zbornik 17. mednarodne multikonference Informacijska družba – IS 2014*. Ljubljana: Inštitut Jožef Stefan.

Kmecl, Matjaž, 2005: *Kratka kulturna zgodovina Slovencev*. Ljubljana: Slovenski PEN.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin in Evan Herbst, 2007: Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. 177–180.

Kohlschütter, Christian, Peter Fankhauser in Wolfgang Nejdl, 2010: Boilerplate Detection Using Shallow Text Features. *WSDM 2010 The Third ACM International Conference on Web Search and Data Mining*. New York: ACM. 441–450.

Kola, Kjersti Wictorsen, 2012: A study of pupils' understanding of the morphological information in the Norwegian electronic dictionary Bokmålsordboka in Nynorskordboka. Vatvedt Fjeld, Ruth in Julie Matilde Torjusen (ur.): *Proceedings of the 15th EURALEX international Congress. EURALEX 2012*. Oslo: Universitetet i Oslo, Institutt for lingvistiske og nordiske studier. 672–675.

Koplenig, Alexander, Peter Meyer in Carolin Müller-Spitzer, 2014: Dictionary users do look up frequent words. A log file analysis. Müller-Spitzer, Carolin (ur.): *Using online dictionaries*. Berlin in Boston: de Gruyter. 229–249.

Korošec, Tomo, 1998: *Stilistika slovenskega poročevalstva*. Ljubljana: Kmečki glas.

Korpus pisnih besedil: specifikacije postopkov za redno zbiranje tekstovnega gradiva za korpus, december 2008. http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik1/SSJ_Kazalnik_1_Specifikacije-pisni-korpus_v1.pdf (dostop 6. 7. 2015).

Kosem, Iztok, 2006: Definicijski jezik v Slovarju slovenskega knjižnega jezika s stališča sodobnih leksikografskih načel. *Jezik in slovstvo* 51/5. 25–45.

Kosem, Iztok, 2010: *Designing a model for a corpus-driven dictionary of academic English.* PhD dissertation. Aston University, UK.

Kosem, Iztok, 2011: Prihodnost leksikografije: dinamični slovar. Jesenšek, Marko (ur.): *Izzivi sodobnega slovenskega slovaropisja*. Maribor: Filozofska fakulteta. 38–48.

Kosem, Iztok, Milos Husak in Diana McCarthy, 2011: GDEX for Slovene. Iztok Kosem in Karmen Kosem (ur.): *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, 10–12 November 2011, Bled, Slovenia*. Ljubljana: Trojina, Institute for Applied Slovene Studies. 151–159.

Kosem, Iztok, 2012: Using GDEX in (semi)-automatic creation of database entries. SKEW-3, 3rd International Sketch Engine workshop, 21–22. marec, Brno, Češka. https://www.sketchengine.co.uk/documentation/attachment/wiki/SKEW-3/Program/GDEX-automatic-entry-extraction-Iztok_Kosem.pdf?format=raw (dostop 9. 8. 2015)

Kosem, Iztok, Mojca Stitar, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman, 2012a: *Analiza jezikovnih težav učencev: korpusni pristop.* Ljubljana: Trojina, zavod za uporabno slovenistiko.

Kosem, Iztok, Polona Gantar in Simon Krek, 2012b: Avtomatsko luščenje leksikalnih podatkov iz korpusa. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012*. Institut Jožef Stefan. 117–122.

Kosem, Iztok in Polona Gantar, 2013: Beleženje in prikazovanje podatkov o jezikovni rabi: od leksikalne baze do spletnega slovarja. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 133–139.

Kosem, Iztok, Polona Gantar in Simon Krek, 2013a: Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0* 1/2. 139–164. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_07.pdf (dostop 8. 8. 2015).

Kosem, Iztok, Polona Gantar in Simon Krek, 2013b: Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. Kosem, Iztok, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets in Maria Tuulik (ur.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana: Trojina, Institute for Applied Slovene Studies in Tallinn: Eesti Keele Instituut. 32–48.

Kosem, Iztok, 2014: Fran: pameten in intuitiven? *Slovenščina 2.0* 2/2. 161–193. http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_09.pdf (dostop 12. 6. 2015).

Kosem, Iztok, 2015a: Oznake: slovarska baza in slovar. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 482–494.

Kosem, Iztok, 2015b: Slovarski zgledi. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 320–339.

Košmrlj, Maja, Jakob Müller, 1972: Pregled tujih kritičnih mnenj in pripomb o Slovarju slovenskega knjižnega jezika. *Jezik in slovstvo* 18/3. 109–112.

Kovačič, Irena (ur.), 1994: *Analiza diskurza*. Ljubljana: Društvo za uporabno jezikoslovje Slovenije.

Kranjc, Simona, 1996/97: Govorjeni diskurz. *Jezik in slovstvo* 42/7. 307–319.

Kranjc, Simona, 1999: *Razvoj govora predšolskih otrok*. Ljubljana: Znanstveni inštitut Filozofske fakultete UL.

Kravos, Nika, 2014: *Primerjava izbranih jezikovnih svetovalnic in preverba skladnosti stališč do posameznih značilnosti svetovalnic v stroki in pri uporabnikih*. Diplomsko delo. Nova Gorica: Univerza v Novi Gorici.

Krek, Simon, 2004: Slovarji serije COBUILD in formalizacija definicijskega jezika. *Jezik in slovstvo* 49/2. 3–16.

Krek, Simon in Adam Kilgarriff, 2006: Slovene Word Sketches. Erjavec. Tomaž in Jerneja Žganec Gros (ur.): *Proceedings of the 5th Slovenian and 1st International Languages Technology Conference.* Ljubljana, Slovenia. 62–67.

Krek, Simon, 2009: Od SSKJ do spletnega portala standardne slovenščine. *Jezik in slovstvo* 54/3-4. 95–113.

Krek, Simon in Tomaž Erjavec, 2009: Standardised encoding of morphological lexica for Slavic languages. Anatoliiovych Shyrokov, Volodymyr in Ludmila Dimitrova (ur.): *MONDILEX Second Open Workshop, Organization and development of digital lexical resources: proceedings*. Kijev: National Academy of Sciences of Ukraine. 24–29.

Krek, Simon, 2010: *Pridobivanje jezikovnih podatkov iz besedilnih korpusov za namen izdelave enojezičnih slovarjev in slovnic.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Krek, Simon, 2011: Language data for digital natives: old wine in a new bottle or...? Plenarno predavanje na konferenci: Electronic lexicography in the 21st century: new applications for new users (eLex2011), Bled, 10–12. november 2011. http://videolectures.net/elex2011_krek_language/?q=simon%20krek (dostop: 27. 7. 2015).

Krek, Simon, 2012a: New Slovene sketch grammar for automatic extraction of lexical data. *SKEW3, tretja mednarodna delavnica orodja Sketch Engine.* Brno, Češka, 21.–22. marec 2012.

Krek, Simon, 2012b: *Slovenski jezik v digitalni dobi/The Slovene Language in the Digital Age.* http://www.meta-net.eu/whitepapers/e-book/slovene.pdf (dostop 6. 7. 2015).

Krek, Simon, 2012c: *Slovenski pravopis – ali je pilot v letalu?* http://www.simonkrek.si/blog/blog_pilot.html (dostop 15. 4. 2013).

Krek, Simon, 2012č: Spletni portal Slogovni priročnik. Krakar Vogel, Boža (ur.): *Slavistika v regijah – Koper. Zbornik Slavističnega društva Slovenije* 23. Ljubljana: Zveza društev Slavistično društvo Slovenije in Znanstvena založba Filozofske Fakultete UL. 225–231.

Krek, Simon, 2013: *Sporazumevanje v slovenskem jeziku: vsebina in rezultati – 2008–2013*. http://videolectures.net/zakljucnakonferencassj2013_krek_vsebina/ (dostop 6. 7. 2015).

Krek, Simon in Iztok Kosem, 2013: *Odgovor na prispevek „SSKJ danes in jutri, potem pa …“*. http://www.sssj.si/datoteke/SSKJ_danes_in_jutri_odgovor.pdf (21. 9. 2013, dostop 15. 1. 2014).

Krek, Simon, Helena Dobrovoljc, Kaja Dobrovoljc in Damjan Popič, 2013a: Online style guide for Slovene as a language resources hub. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference*. Ljubljana: Trojina, Institute for Applied Slovene Studies in Tallinn: Eesti Keele Instituut. 379–391. http://eki.ee/elex2013/proceedings/eLex2013_26_Krek+etal.pdf (dostop 12. 6. 2015).

Krek, Simon, Iztok Kosem in Polona Gantar, 2013b: *Predlog za izdelavo Slovarja sodobnega slovenskega jezika.* Verzija 1.1. http://sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf (dostop 12. 6. 2015).

Krek, Simon, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek, Nanika Holz, 2013c: Training corpus ssj500k 1.3. *Slovenian language resource repository CLARIN.SI.* http://hdl.handle.net/11356/1029 (dostop 30. 6. 2015).

Krek, Simon, 2014a. Prva in druga izdaja SSKJ. *Slovenščina 2.0* 2/2. 114–158. http://www.trojina.org/slovenscina2.0/si/arhiv/2014-2/2014-2-08/ (dostop 21. 6. 2015).

Krek, Simon, 2014b: SSKJ v slovarski bazi. Grahek, Irena in Simona Bergoč (ur.): *Novi slovar za 21. stoletje*. Ljubljana, Ministrstvo za kulturo. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/5-_Simon_Krek_SSKJ_v_slovarski_bazi.pdf (dostop 12. 6. 2015).

Krek, Simon, 2014c: *ZRC SAZU, d. o. o. in SAZU, d. d*. http://www.simonkrek.si/blog/blog_zrc_sazu_doo.html (dostop 27. 7. 2015).

Krek, Simon, 2015a: Slovenska slovnica in računalniško procesiranje besedil v slovenščini. Predavanje na 268. Solomonovem seminarju, Inštitut Jožef Štefan. http://videolectures.net/solomon_krek_slovenska_slovnica (dostop 21. 6. 2015).

Krek, Simon, 2015b: Standardni in knjižni jezik – drugi poskus. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis. Obdobja* 34. Ljubljana: Znanstvena založba Filozofske fakultete UL. V tisku.

Krvina, Domen, 2014: Sprotni slovar slovenskega jezika. Gradivo: okrogla miza Slovensko slovaropisje, Pišece, 2. 10. 2014. *Slavia Centralis* 2. 90–92.

Landau, Sidney I., 1974: Of Matters Lexicographical. Scientific and Technical Entries in American Dictionaries. *American Speech* 49/3-4. 241–244.

Landau, Sidney, I., 1984 (2001): *Dictionaries. The Art and Craft of Lexicography*. New York: Scribners (Cambridge: Cambridge University Press).

Lang, Ewald, 1989: Probleme der Beschreibung von Konjunktionen im allgemeinen einsprachigen Wörterbuch. Hausmann, Franz J., Oskar Reichmann, Herbert E. Wiegand in Ladislav Zgusta (ur.): *Wörterbücher/Dictionaries/Dictionnaires.* Vol. 1. Berlin in New York: de Gruyter. 862–868.

Laufer, Batia, 1993: The Effects of Dictionary Definitions and Examples on the Comprehension of New L2 words. *Cahiers de Lexicologie* 63. 131–142.

Laufer, Batia, 2000: Electronic dictionaries and incidental vocabulary acquisition: does technology make a difference? Heid, Ulrich, Stefan Evert, Egbert Lehmann in Christian Rohrer (ur.): *Proceedings of the Ninth EURALEX International Congress, Stuttgart, Germany, August 8th-12th 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung. 849–854.

Laufer, Batia in Tami Levitzky-Aviad, 2006: Examining the effectiveness of 'Bilingual Dictionary Plus' – a dictionary for production in a foreign language. *International Journal of Lexicography* 19/2. 135–155.

Le dictionnaire de l'Académie française. Marsanne: Redon. CD-ROM z osmimi izdajami slovarja.

Le dictionnaire des enfants, 1991. Pariz: Larousse.

Le dictionnaire des petits ours, 1995. Pariz: Larousse.

Le dictionnaire en herbe, 1989. Pariz: Bordas.

Le grand atelier historique de la langue française. Marsanne: Redon. CD-ROM s 14 starejšimi francoskimi slovarji, med njimi slovarji Nicota (1606), Richeleta (1680), Furetièra (1694) in Férauda (1787).

Lease, Matthew in Omar Alonso, 2014: *Crowdsourcing and Human Computation. Introduction*. Springer.

Ledinek, Nina, 2014a: Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine. Ljubljana: Založba ZRC, ZRC SAZU.

Ledinek, Nina, 2014b: Terminologija v enojezičnem razlagalnem slovarju srednjega obsega. Grahek, Irena in Simona Bergoč (ur.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo*. Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/10-_Nina_Ledinek_-_koncni_prispevek.pdf (dostop 6. 7. 2015)

Leech, Geoffrey, 1992: Corpora and Theories of Linguistic Performance. Svartvik, Jan (ur.): *Directions in Corpus Linguistics*. Berlin in New York: de Gruyter. 105–122.

Leffa, Vilson, 1993: Using an Electronic Dictionary to Understand Foreign Language Texts. *Trabalhos Em Linguistica Aplicada* 21. 19–29.

Legiša, Lino, 1977: *Pisanice 1779–1782*. Ljubljana: Slovenska akademija znanosti in umetnosti.

Lemnitzer, Lothar, 2001: Das Internet als Medium für die Wörterbuchbenutzungsforschung. Lemberg, Ingrid, Bernhard Schröder in Angelika Storrer (ur.): *Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher.* Tübingen. 247–254.

Lemnitzer, Lothar, Christian Pölitz, Jörg Didakowski in Alexander Geyken, 2015: Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German. Kosem, Iztok, Miloš Jakubíček, Jelena Kallas in Simon Krek (ur.): Electronic lexicography in the 21st century: linking lexical data in the digital age. *Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana in Brighton: Trojina, Institute for Applied Slovene Studies in Lexical Computing Ltd. 21–31.

Lenček, Rado L., 1981: *The structure and history of the Slovene language*. Columbus, OH: Slavica Publishers.

Levec, Fran, 1899: *Slovenski pravopis*. Dunaj: Cesarska kraljeva zaloga šolskih knjig.

Levinson, Stephen, 1983: *Pragmatics*. Cambridge: Cambridge University Press.

Lew, Robert, 2002: Questionnaires in dictionary use esearch: a reexamination. Braasch, Anna in Claus Povlsen (ur.): *Proceedings of the Tenth EURALEX International Congress.* Copenhagen: Center for Sprogteknologi. 267–271.

Lew, Robert, 2013: User-generated content (UGC) in online English dictionaries. *OPAL - Online publizierte Arbeiten zur Linguistik*.

Lew, Robert in Gilles-Maurice de Schryver, 2014: Dictionary users in the digital revolution. *International Journal of Lexicography* 27/4. 341–359.

Lew, Robert, 2015: Opportunities and limitations of user studies. Tiberius, Carole in Carolin Müller-Spitzer (ur.): *Research into dictionary* use. *Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie"*. Mannheim: Institut für deutsche Sprache. 6–16.

Lexique actif du français, 2007. Bruxelles: De Boeck.

Liberman, Mark in Alan Prince, 1977: On stress and linguistic rhythm. *Linguistic Inquiry* 8. 249–336.

Linke, Angelika, Markus Nussbaumer in Paul R. Portmann, 2004: *Studienbuch Linguistik.* Tübingen: Max Niemeyer Verlag.

Ljubešić, Nikola in Tomaž Erjavec, 2011: hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. *Text, Speech and Dialogue: Lecture Notes in Computer Science* 6836. 395–402.

Ljubešić, Nikola, Marija Stupar in Terezija Jurić, 2013: Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0* 1/2. 35–57. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_03.pdf (dostop 9. 8. 2015).

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2014: Standardizing tweets with character-level machine translation. Gelbukh, Alexander (ur.): *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014*, *Kathmandu, Nepal*. Heidelberg: Springer. 164–175.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec, 2015: Predicting the level of standardness of text in user-generated content. *Proceedings of the Conference RANLP "Recent Advances in Natural Language Processing".* Hissar, Bolgarija.

Logar, Nataša, 2000/2001: Kvalifikator ekspr. v Slovarju slovenskega knjižnega jezika na ravni frazeologije. *Jezik in slovstvo* 46/4. 137–148.

Logar, Nataša, 2006: Stilno zaznamovane nove tvorjenke: tipologija. *Slavistična revija* 54/pos. št. 87–101.

Logar Berginc, Nataša in Špela Vintar, 2008: Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izrazja. *Jezik in slovstvo* 53/5. 3–17.

Logar Berginc, Nataša, 2009: Slovenski splošni in terminološki slovarji: za koga? Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 225–231.

Logar Berginc, Nataša in Simon Šuster, 2009: Gradnja novega korpusa slovenščine. *Jezik in slovstvo* 54/3-4. 57–68.

Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek, 2012a: *Korpusi slovenskega jezika Gigafida, Kres, ccGigafida in ccKRES: gradnja, vsebina, uporaba*, Ljubljana: Trojina, zavod za uporabno slovenistiko in Fakulteta za družbene vede.

Logar, Nataša, Špela Vintar in Špela Arhar Holdt, 2012b: Luščenje terminoloških kandidatov za slovar odnosov z javnostmi. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.). *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 135–140.

Logar, Nataša, 2013: *Korpusna terminografija: primer odnosov z javnostmi*. Ljubljana: Trojina, zavod za uporabno slovenistiko in Fakulteta za družbene vede.

Logar Berginc, Nataša in Nikola Ljubešić, 2013: Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0* 1/1. 78–110. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_05.pdf (dostop 10. 8. 2015).

Logar Berginc, Nataša, Špela Vintar in Špela Arhar Holdt, 2013: Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka. *Slovenščina 2.0* 1/2. 113–138. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_06.pdf (dostop 10. 8. 2015).

Logar, Nataša, 2014: Verodostojnost korpusa kot gradivnega vira za slovar. Grahek, Irena in Simona Bergoč (ur.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo*. Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/7-_Natasa_Logar_-prispevek_-za_oddajo.pdf http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/20-_Tomaz_Erjavec-SlovarPosvet.pdf (dostop 6. 7. 2015).

Logar, Nataša, 2015a: Gradnja referenčnih korpusov na novo: nadgradnja Gigafide. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 218–241.

Logar, Nataša, Nikola Ljubešić in Tomaž Erjavec, 2015b: KRES in Gigafida kot korpusna osnova za slovar: podobnosti in razlike. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis. Obdobja* 34. Ljubljana: Znanstvena založba Filozofske fakultete UL. V tisku.

Lorentzen, Henrik in Liisa Theilgaard, 2012: Online dictionaries – how do users find them and what do they do once they have? Vatvedt Fjeld, Ruth in Julie Matilde Torjusen (ur.)*: Proceedings of the 15th EURALEX International Congress*. *EURALEX 2012*. Oslo: Universitetet i Oslo, Institutt for lingvistiske og nordiske studier. 654–660.

Louw, Bill, 1993: Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. Baker, Mona, Gill Francis in Elena Tognini-Bonelli (ur.): *Text and technology. In honour of John Sinclair*. Amsterdam in Philadelphia: John Benjamins. 157–176.

Louw, Bill, 2000: Contextual Prosodic Theory: bringing semantic prosodies to life. Heffer, Cris in Helen Sauntson (ur.): *Words in Context: A tribute to John Sinclair on his retirement*. Birmingham: University of Birmingham. 48–94.

Lui, Marco in Timothy Baldwin, 2012: langid. py: an Off-the-Shelf Language Identification Tool. *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistic. 25–30. http://www.aclweb.org/anthology/P12-3005 (dostop 15. 4. 2015).

Mackintosh, Kristen, 1998: An empirical study of dictionary use in L2-L1 translation. Atkins, B. T. Sue (ur.): *Using Dictionaries*. Tübingen: Niemeyer. 123–149.

Majcenovič, Helena, 1999: Praktični vidiki normiranja v slovarjih. *Jezikoslovni zapiski* 5. 63–90.

Majcenovič, Helena, 2001: *Spremembe v pravopisnih pravilih: razlike med 5. in 6. izdajo: (5. izdaja: Slovenski pravopis 1 - Pravila, 1997; 6. izdaja: Slovenski pravopis 2001)*. http://www.zrc-sazu.si/secret_portals/pravopis/pdf/spremembe.pdf (dostop 15. 7. 2015).

Malmgren, Sven-Göran, 2009: On production-oriented information in Swedish monolingual defining dictionaries. Tarp, Sven, Sandro Nielsen in Henning Bergenholtz (ur.): *Lexicography in the 21st Century: In Honour of Henning Bergenholtz*. Amsterdam in Philadelphia: John Benjamins. 93–102.

Marcus, Solomon, 1970: Définitions logiques et définitions lexicographiques. *Langages* 19. 87–92.

Martin, J. R., 2003: Cohesion and texture. Schiffrin, Deborah, Deborah Tannen in Heidi E. Hamilton (ur.): *The Handbook of Discourse Analysis*. Oxford: Blackwell. 35–53.

Martin, Robert, 1990: La définition »naturelle«. Chaurand, Jacques in Francine Mazière (ur.): *La définition*. Pariz: Larousse. 86–95.

Martin, Robert, 1991: Typicité et sens de mots. Dubois, Danièle (ur.): *Sémantique et cognition. Catégories, prototypes, typicalité*. Pariz: CNRS. 151–159.

Martinez, Ignacio M. Palacios, 2011: I might, I might go I mean it depends on money things and stuff: A preliminary analysis of general extenders in British teenagers' discourse. *Journal of Pragmatics* 43/9. 2452–2470.

Marušič, Franc in Rok Žaucer, 2007: O določnem ta v pogovorni slovenščini (z navezavo na določno obliko pridevnika). *Slavistična revija* 55/1-2. 223–247.

McCreary, Don R. in Fredric T. Dolezal, 1999: A Study of Dictionary Use by ESL Students in an American University. *International Journal of Lexicography* 12/2. 107–146.

McCreary, Don R., 2002: American Freshmen and English Dictionaries: ‚I Had Aspersions of Becoming an English Teacher'. *International Journal of Lexicography* 15/3. 181–205.

McCreary, Don R., 2004: Labelling of Pejorative Terms in a Dictionary of College Slang. Williams, Geoffrey in Sandra Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress. EURALEX 2004*. Lorient: Universite de Bretagne-Sud. 891–896.

McDonald, Ryan, Kevin Lerman in Fernando Pereira, 2006: Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. *Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, *NYC, USA.* Stroudsburg, ZDA: Association for Computational Linguistics. 216–220.

McEnery, Tony in Andrew Hardie, 2012: *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mentrup, Wolfgang, 1984: Wörterbuchbenutzungssituationen–Sprachbenutzungssituationen. Anmerkungen zur Verwendung einiger Termini bei HE Wiegand. Besch, Werner, Klaus Hufeland, Volker Schupp in Peter Wiehl (ur.): *Festschrift f*ür Siegfried Grosse zum 60. Geburtstag. Göppingen: Kümmerle Verlag. 143–173.

Meschonnic, Henri, 1991: *Des mots et des mondes: dictionnaires, encyclopâedies, grammaires, nomenclatures*. Paris: Hatier.

Meyer, Christian in Iryna Gurevych, 2012: Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. *Electronic Lexicography.* 259–291.

Migla, Ilga in Ieva Zuicena, 2014: The Dictionary of Contemporary Latvian Language and its Lexicographical Process. *Workflow of Corpus-based Lexicography, COST ENeL WG3 meeting, Bolzano, 19. julij*. http://www.elexicography.eu/wp-content/uploads/2014/07/Migla_2014_COST_Bolzano.pdf (dostop 6. 7. 2015).

Mikolič, Vesna, 2007: Modifikacija podstave in argumentacijska struktura besedilnih vrst. *Slavistična revija* 55/1-2. 341–355.

Mikolič, Vesna, 2013: Področni govor in terminologija na primeru jezika turizma. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 255–261.

Mikolič, Vesna, 2015: Slovarski uporabniki – ustvarjalci: ustvarjati v jeziku in z jezikom. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 182–195.

Mikolič, Vesna in Maša Rolih, 2015: Besedilna zvrstnost v novih medijih kot slovarska vsebina. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis. Obdobja* 34. Ljubljana: Znanstvena založba Filozofske fakultete UL. V tisku.

Miller, George in Patricia Gildea, 1987: How Children Learn Words. *Scientific American* 257/3. 94–99.

Milroy, James in Lesley Milroy, 1998 (1999): *Authority in language: Investigating Standard English*. London in New York: Routledge.

Milroy, James, 2001: Language ideologies and the consequences of standardization. *Journal of Sociolinguistics* 4/5. 530–555.

Mistrík, Jozef, 1985: Štylistika. Bratislava: Slovenské pedagogické nakladatelstvo.

Mitchell, Evelyn, 1983: *Search-Do Reading: Difficulties in Using a Dictionary.* Aberdeen: College of Education.

Moon, Rosamund, 1987: The Analysis of Meanning. Sincair, John M. (ur.): *Looking up: An account of the COBUILD Project in Lexical Computing*. London in Glasgow: Collins ELT. 86–103.

Moon, Rosamund, 1998: On using spoken data in corpus lexicography. Fontenelle, Thierry, Philippe Hiligsmann, Archibald Michiels, André Moulin in Siegfried Theissen (ur.): *Proceedings* of the Eighth *EURALEX* International Congress. *EURALEX 1998.* Liège: Université de Liège, Département d'anglais et de néerlandais. 347–355.

Morley, G. David, 2000: *Syntax in Functional Grammar.* London, New York: Continuum.

Motschenbacher, Heiko, 2010: *Language, Gender and Sexual Identity: Poststructuralist Perspectives*. Amsterdam in Philadelphia: John Benjamins.

Müller-Spitzer, Carolin, Alexander Koplenig in Antje Töpel, 2011: What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. Kosem, Karmen in Iztok Kosem (ur): *Proceedings of eLex 2011, Bled, 10–12 November 2011: Electronic Lexicography in the 21st Century - New Applications for New Users.* Ljubljana: Trojina, Institute for Applied Slovene Studies. 203–208.

Müller-Spitzer, Carolin (ur.), 2014: *Using Online Dictionaries*. Berlin in Boston: de Gruyter.

Müller-Spitzer, Carolin, Sascha Wolfer in Alexander Koplening, 2015: Observing Online Dictionary Users: Studies Using Wiktionary Log Files. *International Journal of Lexicography* 28/1. 1–26.

Müller, Jakob, 1982: Jezik kot vrednota ali jezik kot resničnost? *Naši razgledi* 10 (28. maj 1982). 294–296.

Müller, Jakob, 1996: Slovar slovenskega knjižnega jezika in kritika z bibliografijo (1960–1992). *Razprave SAZU. Razred za filološke in literarne vede* 15. Ljubljana: SAZU. 187–234.

Müller, Jakob, 2009: Kritične misli in zamisli o SSKJ. Perdih, Andrej (ur.): *Strokovni posvet o slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 17–21, 25.

Negri, Matteo, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo in Alessandro Marchetti, 2011: Divide and conquer: crowdsourcing the creation of crosslingual textual entailment corpora. *Conference on Empirical Methods in Natural Language Processing, EMNLP '11. Proceedings of the Conference.* Stroudsburg, ZDA: Association for Computational Linguistics. 670–679.

Nesi, Hilary, 1996: The Role of Illustrative Examples in Productive Dictionary Use. *Dictionaries* 17. 198–206.

Nesi, Hilary, 2000: *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries*. Tübingen: Max Niemeyer Verlag.

Nesi, Hilary in Richard Haill, 2002: A Study of Dictionary Use by International Students at a British University. *International Journal of Lexicography* 15/4. 277–305.

Nesi, Hilary, 2011: The effect of e-dictionary font on vocabulary retention. *Electronic lexicography in the 21st century: new applications for new users (eLex), 10–12 November 2011, Bled, Slovenia*. http://videolectures.net/elex2011_nesi_effect/ (dostop 9. 8. 2015).

Nespor, Marina in Irene Vogel, 1986: *Prosodic phonology*. Dordrecht: Foris.

Neubach, Abigail in Andrew Cohen, 1988: Processing Strategies and Problems Encountered in the Use of Dictionaries. *Dictionaries: Journal of the Dictionary Society of North America* 10. 1–19.

Neubauer, Fritz, 1987: How to Define a Defining Vocabulary. Ilson, Robert F.: *A Spectrum of Lexicography*. Amsterdam in Philadephia: John Benjamims. 49–59.

Neustupný, Jiři V., 1968: Some general aspects of "language" problems and "language" policy in developing societies. Fishman, Joshua A., Charles A. Ferguson in Jyotirindra das Gupta (ur.): *Language problems of the developing nations*. New York, London, Sydney in Toronto: John Wiley & Sons, Inc. 285–294.

Neustupný, Jiři V., 1978: *Post-structural approaches to language*. Tokyo: University of Tokyo Press.

Newman, Andrew, 2007: *A Relational View of the Semantic Web*. http://www.xml.com/pub/a/2007/03/14/a-relational-view-of-the-semantic-web.html (dostop 21. 6. 2015).

Nidorfer Šiškovič, Mojca, 2013: Žanrskost funkcijskih besedilnih vrst. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 269–275.

Nivre, Joakim et al., 2105: *Universal Dependencies 1.0*. http://hdl.handle.net/11234/1-1464 (dostop 30. 6. 2015).

Norgaard, Nina, Rocio Montoro in Beatrix Busse, 2010: *Key terms in stylistics*. New York: Continuum.

Nouveau Petit Robert, 2014. Pariz: Le Robert.

Novak, France, 1963: Ob poskusnem snopiču Slovarja slovenskega knjižnega jezika. *Sodobnost*. 467–470.

Nuccorini, Stefania, 1992: Monitoring Dictionary Use. Tommola, Hannu, Krista Varantola, Tarja Salmi-Tolonen in Jurgen Schopp (ur.): *EURALEX ,92 Proceedings I-II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland. Studia translatologica.* 89–102.

Oblak, Tanja, Gregor Petrič, Marko Pahor, Franc Trček in Slavko Splichal, 2005: *Splet kot medij in mediji na spletu.* Ljubljana: Fakulteta za družbene vede.

OECD Principles and Guidelines for Access to Research Data from Public Funding. http://www.oecd.org/sti/sci-tech/38500813.pdf (dostop 6. 7. 2015).

Ogrin, Matija, Jan Jona Javoršek in Tomaž Erjavec, 2013: A register of early modern Slovenian manuscripts. *Journal of the Text Encoding Initiative* 4. 1–13. http://jtei.revues.org/715 (dostop 15. 1. 2015).

Open Access Slovenia. http://www.openaccess.si/ (dostop 6. 7. 2015).

Osswald, Rainer: Syntax and Lexicography. Alexiadou, Artemis in Tibor Kiss (ur.): *Syntax II. An International Handbook. Handbooks of Linguistics and Communication Science*. Berlin: de Gruyter. V tisku.

Oyama, Satoshi, Yukino Baba, Yuko Sakurai in Hisashi Kashima, 2013: Accurate Integration of Crowdsourced Labels Using Workers' Self-Reported Confidence Scores. *Proceedings* of the *Twenty-Third International Joint Conference on Artificial Intelligence*. 2554–2560.

Partington, Alan S., 1998: Patterns and Meanings. Using Corpora for English Language Research and Teaching. *Studies in Corpus Linguistics* 2. Amsterdam in Philadelphia: John Benjamins.

Partington, Alan S., 2004: 'Utterly content in each other's company' Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9. 131–156.

Paulsen Christensen, Tina, 2011: Studies on the mental processes in translation memory assisted translation – the State of the Art. *trans-kom* 2/4. 137–160.

Pavelec, Daniel, Luiz S. Oliveira, Edson J. Justino in Leonardo Batista, 2008: Using Conjunctions and Adverbs for Author Verification. *Journal of Universal Computer Science*. 14/18. 2967–2981.

Paynter, Diane E., Elena Bodrova in Jane K. Doty, 2005: *For the Love of Words: Vocabulary Instruction that Works*. San Francisco: Jossey-Bass teacher.

Pečjak, Sonja, 2012: *Psihološki vidiki bralne pismenosti. Od teorije k praksi.* Ljubljana: Znanstvena založba Filozofske fakultete.

Peperkamp, Sharon, 1997: *Prosodic words*. The Hague: Holland Academic Graphics.

Perdih, Andrej, 2009 (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU.

Petek, Bojan, Rastislav Šuštaršič in Smiljana Komar, 1996: An acoustic analysis of contemporary vowels of the standard Slovenian language. *Proceedings ICSLP 96: Fourth International Conference on Spoken Language Processing, October 3–6, 1996*. Philadelphia, PA, USA, Newark, DE. 133–136. http://www.asel.udel.edu/icslp/cdrom/vol1/820/a820.pdf (dostop 11. 8. 2015).

Peters, Pam in Trinidad Fernandez, 2013: The lexical needs of ESP students in a professional field. *English for Specific Purposes* 32/4. 236–247.

Petit Robert des enfants, 1988. Pariz: Le Robert, 1988.

Petrylaitė, Regina, Diana Vaškelienė in Tatjana Věžytė, 2008: Changing Skills of Dictionary Use. *Studies about Languages* 12. 77–82.

Philip, Gill, 2009: Why prosodies aren't always there: Insights into the idiom principle. Mahlberg, Michaela, Victorina González-Díaz in Catherine Smith (ur.): *Proceedings of the Corpus Linguistics Conference CL2009*. Liverpool: University of Liverpool. http://ucrel.lancs.ac.uk/publications/cl2009 (dostop 17. 10. 2011).

Pierrehumbert, Janet M., 2003: Word-specific phonetics. Gussenhoven, Carlos in Natasha Warner (ur.): *Frequent words are more subject to lenition than less frequent words. Laboratory Phonology* VII. Berlin in New York: de Gruyter. 101–140.

Pisanski Peterlin, Agnes, 2010: Se za strukturiranje besedila v prevodih uporabljajo drugačni elementi kot v izvirnikih? Korpusna analiza medpovednega in medstavčnega *in*. Vintar, Špela (ur.): *Slovenske korpusne raziskave.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 156–177.

Pisanski Peterlin, Agnes, 2011: *Metabesedilo med dvema kulturama*. Ljubljana: Znanstvena založba Filozofske fakultete UL.

Pisanski Peterlin, Agnes, 2015: Problematika veznika kot besedne vrste v enojezičnem slovarju. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 514–523.

Pogorelec, Breda, 1964a: Ob poskusnem snopiču Slovarja slovenskega knjižnega jezika. *Jezik in slovstvo* 9/7-8. 232–242.

Pogorelec, Breda, 1964b: *Veznik v slovenščini.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Pollak, Senja, 2014: *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Pomikálek, Jan, 2011: *jusText.* LINDAT/CLARIN Digital Library at Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11858/00-097C-0000-000D-F696-9 (dostop 22. 5. 2015).

Popič, Damjan, 2014: *Korpusnojezikoslovna analiza vplivov na slovenska prevodna besedila*. Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Popič, Damjan in Vojko Gorjanc, 2014: Prevodna dejavnost v jezikovni politiki in jezikovnem načrtovanju: od nacionalnega k nadnacionalnemu. *Teorija in praksa* 51/4. 583–599.

Popič, Damjan, 2015: Normativna informacija v sodobnem slovarju. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL. 106–133.

Predmetnik osnovne šole, 2014. Ljubljana Ministrstvo za izobraževanje, znanost in šport RS. http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/podrocje/os/devetletka/predmetniki/Pred_14_OS_4_12.pdf (dostop 6. 7. 2015).

Priporočilo Komisije z dne 17. julija 2012 o dostopu do znanstvenih informacij in njihovem arhiviranju.http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:SL:PDF (dostop 6. 7. 2015).

Prunč, Erich, 2009: Veliki čudež malega jezika. *Jezik in slovstvo* 54/1. 5–12.

Pruvost, Jean, 2001: Les dictionnaires d'apprentissage monolingues de langue française: problèmes et méthodes. Pruvost, Jean: *Les dictionnaires de langue fran*çaise. Pariz: Honoré Champion. 67–95.

Pruvost, Jean, 2003: Les dictionnaires français monolingues d'apprentissage: une histoire récente et renouvelée. *Quaderni del CIRSIL* 2. 23–56.

Pustejovsky, James, 1995: *The Generative Lexicon.* Cambridge, Massachusetts: The MIT Press.

Quemada, Bernard, 1987: Notes sur la lexicographie et dictionnairique. *Cahiers de lexicologie* 51/2. 229–242.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech in Jan Svartvik, 1985 (1992): *A Comprehensive Grammar of the English Language*. London in New York: Longman.

Ramm, Wiebke in Cathrine Fabricius-Hansen, 2005: Coordination and discourse-structural salience from a cross-linguistic perspective. Stede, Manfred, Christian Chiarcos, Michael Grabski in Luuk Lagerwerf (ur.): *Salience in Discourse: Multidisciplinary Approaches to Discourse 2005.* Münster: Stichting in Nodus Publ. 119–128.

Rayson, Paul in Roger Garside, 2000: Comparing Corpora Using Frequency Profiling. *Proceedings of the ACL Workshop on Comparing Corpora*. Hong Kong. 1–6. http://www.comp.lancs.ac.uk/~rayson/publications/rg_acl2000.pdf (dostop 22. 3. 2014).

Rayson, Paul, Andrew Wilson in Geoffrey Leech, 2001: Grammatical word class variation within the British National Corpus sampler. *Language and Computers* 36/1. 295–306.

Reffle, Ulrich, 2011: Efficiently generating correction suggestions for garbled tokens of historical language. *Natural Language Engineering* 17. 265–282.

Rey-Debove, Josette, 1971: Étude linguistique et sémiotique des dictionnaires français contemporain. Hag in Pariz: Mouton.

Rey-Debove, Josette, 1989: Prototypes et définitions. *DRLAV* 41. 143–167.

Rey-Debove, Josette, 1991: La lexicographie moderne. *Travaux de linguistique* 23. 145–159.

Rey, Alain, 1977: *Le lexique: images et modèles – du dictionnaire* à la lexicologie. Pariz: Armand Colin.

Rey, Alain, 1990: Definitional Semantics: Its Evolution in French Lexicography. Tomaszczyk, Jerzy in Barbara Lewandowska-Tomaszczyk: *Meaning and Lexicography*. Amsterdam in Philadelphia: John Benjamins. 43–55.

Rigler, Jakob, 1971: H kritikam pravopisa, pravorečja in oblikoslovja v SSKJ. *Slavistična revija* 19/4. 433–462.

Rigler, Jakob, 1972: H kritikam pravopisa, pravorečja in oblikoslovja v SSKJ. *Slavistična revija* 20/1. 244–251.

RIS: Raba interneta v Sloveniji. http://www.ris.org/ (dostop 12. 6. 2015).

Roberts, Roda P., 1992: Translation pedagogy: strategies for improving dictionary use. *TTR: traduction, terminologie, rédaction* 1/5. 49–76.

Roberts, Rona P., 1997: Using dictionaries efficiently. *38th Annual Conference of the American Translators Association*. San Francisco, CA. http://www.dico.uottawa.ca/articles-en.htm (dostop: 15. 6. 2015).

Rojc, Matej, Zdravko Kačič in Darinka Verdonik, 2002: Design and implementation of the Slovenian phonetic and morphology lexicons for the use in spoken language applications. *Proceeding od the Third international conference on language resources and evaluation, Las Palmas de Grand Canaria*. Grand Canaria: European Language Resources Association. 1296–1300.

Romih, Miro, 1998: Direktorijska struktura korpusa FIDA. *Uporabno jezikoslovje* 6. 79–84.

Romih, Miro in Peter Holozan, 2002: Slovensko-angleški prevajalni sistem. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Tretje konference Jezikovne tehnologije*. Institut Jožef Stefan. 167.

Romih, Miro in Simon Krek, 2012: Termania – prosto dostopni spletni slovarski portal. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 163–166.

Rozman, Tadeja, 2003: *Tipi slovarjev glede na naslovnika (Pragmatične informacije v enojezičnih slovarjih za tujce).* Diplomsko delo. Ljubljana: Filozofska fakulteta UL.

Rozman, Tadeja, 2009: The Dictionary of Standard Slovenian – A(n) (Un)faithful Companion? Granić, Jagoda (ur.): *Jezična politika i jezična stvarnost/Language Policy and Language Reality*. Zagreb: HDPL. 126–136.

Rozman, Tadeja, 2010: *Vloga enojezičnega razlagalnega slovarja slovenščine pri razvoju jezikovne zmožnosti.* Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Rozman, Tadeja, Irena Krapš Vodopivec, Mojca Stritar, Iztok Kosem in Simon Krek, 2010: *Nova didaktika poučevanja slovenskega jezika – projekt »Sporazumevanje v slovenskem jeziku« – Kazalnik 15*. http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K15.aspx (dostop 12. 6. 2015).

Rozman, Tadeja, Irena Krapš Vodopivec, Mojca Stritar in Iztok Kosem, 2012: *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Rozman, Tadeja, Iztok Kosem, Nataša Pirih Svetina in Ina Ferbežar, 2015: Slovarji in učenje slovenščine. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 150–167.

Rumshisky, Anna, 2011: Crowdsourcing Word Sense Definition. *LAW* V. *Fifth* Linguitic Annotation Workshop. *Proceedings of the Workshop. Portland, ZDA: Association for Computational Linguistics.* 74–81.

Rumshisky, Anna, Nick Botchan, Sophie Kushkuley in James Pustejovsky, 2012: Word Sense Inventories by Non-Experts. *Proceedings* of the Eighth International Conference on Language Resources and Evaluation. LREC '12. Istanbul, Turky.

Rundell, Michael, 2006: More than one way to skin a cat: why full-sentence definitions have not been universally adopted. Corino, Elisa, Carla Marello in Cristina Onesti (ur.): *EURALEX 2006 Proceedings*. Torino: EURALEX. 323–337.

Rundell, Michael in Adam Kilgarriff, 2011: Automating the creation of dictionaries: where will it all end? Meunier, Fanny, Sylvie De Cock, Gaëtanelle Gilquin in Magali Paquot (ur.): *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam in Philadelphia: John Benjamins. 257–281.

Rundell, Michael in Sue Atkins, 2011: The DANTE database: a User Guide. Kosem, Iztok in Karmen Kosem (ur.): *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011, 10-12 November 2011, Bled, Slovenia*. Ljubljana: Trojina, Institute for Applied Slovene Studies. 233–246.

Rundell, Michael, 2014: Macmillan English Dictionary: The End of Print? *Slovenščina 2.0* 2/2. 1–14. http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_02.pdf (dostop 17. 7. 2015).

Rupnik, Jan, Miha Grčar in Tomaž Erjavec, 2010: Improving Morphosyntactic Tagging of Slovene Language Through Meta-tagging. *Informatica* 34/2. 169–175.

Rychly, Pavel, 2007: Manatee/bonito - a modular corpus manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Univerza Masaryk. 65–70.

Sabou, Marta, Kalina Bontcheva, Leon Derczynski in Arno Scharl, 2014: Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC '14.* Reykjavik, Iceland. 859–866.

Salton, Gerard in Christopher Buckley, 1988: Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management: an International Journal* 24/5. 513–523.

Salvi, Giampaolo, 2013: *Le parti del discorso.* Roma: Carocci.

Sanchez Ramos, Maria del Mar, 2005: Research on Dictionary Use by Trainee Translators. *Translation Journal* 2/9. http://www.translationdirectory.com/article476.htm (dostop 15. 6. 2015).

Scherrer, Yves in Tomaž Erjavec, 2013: Modernizing historical Slovene words with character-based SMT. *The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. 58–62. http://hal.archives-ouvertes.fr/docs/00/83/85/75/PDF/13-scherrer-modernize.pdf (dostop 15. 1. 2015).

Schiffrin, Deborah, 1987: *Discourse Markers.* Cambridge: Cambridge University Press.

Schlamberger Brezar, Mojca, 2009: *Povezovalci v francoščini: od teoretičnih izhodišč do analize v diskurzu.* Ljubljana: Znanstvena založba Filozofske fakultete UL.

Schryver, Gilles-Maurice de, 2003: Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16/2. 143–199.

Schutz, Rik, 2002: Indirect Offensive Language in Dictionaries. Braasch, Anna in Claus Povlsen (ur.): *Proceedings of the 10th EURALEX International Congress*. Copenhagen: Center for Sprogteknologi. 637–641.

Scott, Mike, 1997: PC analysis of key words – and key key words. *System* 25/1. 1–13.

Scott, Mike, 1998: Focusing on the Text and its Key Words. Stephens, C. (ur.): *TALC ,98 Proceedings*. Oxford: Humanities Computing Unit, Oxford University. 152–164.

Selkirk, Elisabeth, 1980: The role of prosodic categories in English word stress. *Linguistic Inquiry* 11. 563–606.

Selva, Thierry in Serge Verlinde, 2002: L'utilisation d'un dictionnaire électronique: une étude de cas. Anna Braasch in Claus Povlsen (ur.): *Proceedings of the 10th EURALEX International Congress*. Copenhagen: Center for Sprogteknologi. 773–781.

Sharoff, Serge, 2010: Analysing Similarities and Differences between Corpora. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Sedme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 5–11.

Siepmann, Dirk, 2015: Dictionaries and spoken language: A corpus-based review of French dictionaries. *International Journal of Lexicography* 28/2. 139–168.

Siepmann, Dirk in Christoph Bürgel, 2015: L'élaboration d'une grammaire pédagogique à partir de corpus: l'exemple du subjonctif. Tinnefeld, Thomas (ur.): *Grammatikographie und Didaktische Grammatik – gestern, heute, morgen*. Saarbrücken: htw saar.

Silberman, M. Six, Joel Ross, Lilly Irani in Bill Tomlinson, 2010: Sellers' problems in human computation markets. *Proceedings of the ACM SIGKDD Workshop on Human Computation*. 18–21.

Silić, Josip in Ivo Pranjković, 2005: *Gramatika hrvatskoga jezika.* Zagreb: Školska knjiga.

Sinclair, John McH., 1987: *Looking Up: An Account of the COBUILD Project in Lexical Computing.* London: Collins.

Sinclair, John McH. in Patrick Hanks (ur.), 1987: *Collins COBUILD English Language Dictionary*. London in Glasgow: Collins.

Sinclair, John McH., 1991: *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John McH., 1996: The Search for Units of Meaning. *Textus* 9. 75–106.
Sinclair, John McH., 1999: A way with common words. Hasselgård, Hilde in Signe Oksefjell (ur.): *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam in Atlanta, GA: Rodopi.
Sinclair, John McH., 2004: *Trust the Text*. London in New York: Routledge.
Singleton, David, 1999: *Exploring the Second Language Mental Lexicon.* Cambridge: Cambridge University Press.
Skubic, Andrej, 1995: Klasifikacija funkcijske zvrstnosti in pragmatična definicija funkcije. *Jezik in slovstvo* 40/5. 155–168.
Skubic, Andrej, 1999: Ogled kohezijske vloge slovenskega členka. *Slavistična revija* 47/2. 211–238.
Skubic, Andrej E., 2003: Mesto standardnega jezika v jezikovnem repertoarju posameznika. Ada Vidovič Muha (ur.): *Slovenski knjižni jezik – aktualna vprašanja in zgodovinske izkušnje. Obdobja* 20. Ljubljan: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete UL. 209–226.
Skubic, Andrej E., 2004: Sociolekti od izraza do pomena: kultiviranost, obrobje in ekscesi. Kržišnik, Erika (ur.): *Aktualizacija jezikovnozvrstne teorije na Slovenskem: členitev jezikovne resničnosti. Obdobja* 22. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete UL. 297–320.
Skubic, Andrej E., 2005: *Obrazi jezika*. Ljubljana: Študentska založba.
Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje. Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva, 2011. http://www.europass.si/files/userfiles/europass/SEJO%20komplet%20za%20splet.pdf (dostop 22. 6. 2015).
Slovar novejšega besedja slovenskega jezika, 2012. Ur. Aleksandra Bizjak Končar in Marko Snoj. Ljubljana: Založba ZRC, ZRC SAZU.
Slovar slovenskega knjižnega jezika, 1970–1991. Ljubljana: DZS in Slovenska akademija znanosti in umetnosti, Inštitut za slovenski jezik.
Slovenski pravopis, 2001. Gl. ur. Jože Toporišič. Ljubljana: Založba ZRC, ZRC SAZU.
Smolej, Mojca, 2001: *Členek v slovenskem knjižnem jeziku: pomenoslovni in skladenjski vidiki*. Magistrsko delo. Ljubljana: Filozofska fakulteta UL.
Smolej, Mojca, 2004a: Členki kot besedilni povezovalci. *Jezik in slovstvo* 49/5. 45–57.
Smolej, Mojca, 2004b: Obvezni in neobvezni členki. *Slavistična revija* 52/2. 141–155.
Smolej, Mojca, 2009: Češki vplivi na Toporišičevo teorijo členkov. *Opera Slavica* 19/2. 11–22.
Snoj, Jerica, 2010: *Metafora v leksikalnem sistemu*. Ljubljana: Založba ZRC, ZRC SAZU.
Snow, Rion, Brendan O'Connor, Daniel Jurafsky in Andrew Y. Ng, 2008: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Conference on Empirical Methods in Natural Language Processing, EMNLP '11. Proceedings of the Conference.* Stroudsburg, ZDA: Association for Computational Linguistics. 254–263.
Sorlin, Sandrin, 2014: The ‚interdisciplinarity' of stylistics. *Topics in Linguistics* 14/1. 9–15.
»Sporazumevanje v slovenskem jeziku«. http://www.slovenscina.eu/ (dostop 24. 6. 2015).
Sproat, Richard, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf in Christofer Richards, 2001: Normalization of Non-Standard Words. *Computer Speech and Language* 15/3. 287–333.

Srebot Rejec, Tatjana, 1988: Word accent and vowel duration in Standard Slovene: An acoustic and linguistic investigation. *Slavistische Beiträge* 226. München: Otto Sagner.

Srebot Rejec, Tatjana, 1998: O slovenskih samoglasniških sestavih zadnjih 45 let. *Slavistična revija* 46/4. 339–346.

Stabej, Jože, 1977: *Hieronymus Megiser. Thesaurus polyglottus.* Iz njega je slovensko besedje z latinskimi in nemškimi pomeni za slovensko-latinsko-nemški slovar izpisal in uredil Jože Stabéj. Ljubljana: Slovenska akademija znanosti in umetnosti.

Stabej, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6. 96–106.

Stabej, Marko in Primož Vitez, 2000: KGB (korpus govorjenih besedil) v slovenščini. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 79–81.

Stabej, Marko, Tadeja Rozman, Nataša Pirih Svetina, Nina Modrijan in Boštjan Bajec, 2008: *Jezikovni viri pri jezikovnem pouku v osnovni in srednji šoli: končno poročilo z rezultati dela*. Ljubljana: Pedagoški inštitut. http://www.trojina.si/p/jezikovni-viri-pri-jezikovnem-pouku-v-osnovni-in-srednji-soli/ (dostop 22. 6. 2015).

Stabej, Marko, 2009a (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Stabej, Marko, 2009b: Slovarji in govorci: kot pes in mačka? *Jezik in slovstvo,* 54/3-4. 115–138.

Stabej, Marko, 2010: *V družbi z jezikom*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Stabej, Marko, 2015a: Daj mi slovar in spremenim ti (jezikovno) skupnost. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 16–31.

Stabej, Marko, 2015b: L'État, ce n'est pas moi. Tivadar, Hotimir (ur.): *Država in narod v slovenskem jeziku, literaturi in kulturi*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 27–36.

Stabej, Marko, 2015c: Nekaj misli o označevanju in stilistiki v »novem slovarju«. Interno gradivo za stilistično skupino. 23. 1. 2015.

Steele, James, 1990: *Meaninh-Text Theory: Linguistics, Lexicography, and Implications*. Ottawa: University of Ottawa Press.

Sterkenburg, Piet van (ur.), 2003: *A Practical Guide to Lexicography*. Amsterdam in Philadelphia: John Benjamins.

Stewart, Dominic, 2010: *Semantic prosody: a critical evaluation*. London in New York: Routledge.

Stritar, Mojca, 2012: *Korpusi usvajanja tujega jezika.* Ljubljana: Zveza društev Slavistično društvo Slovenije.

Stubbs, Michael, 1995a: Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language* 2/1. 23–55.

Stubbs, Michael, 1995b: Corpus evidence for norms of lexical collocation. Cook, Guy in Barbara Seidlhofer (ur.): *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press. 245–56.

Stubbs, Michael, 2001: On inference theories and code theories: corpus evidence for semantic schemas. *Text* 21/3. 436–465.

Suchomel, Vít in Jan Pomikálek, 2012: Efficient Web Crawling for Large Text Corpora. Kilgarriff, Adam in Serge Sharoff (ur.): *Proceedings of the 7th Web as Corpus Workshop (WAC7)*. Lyon. 39–43. http://nlp.fi.muni.cz/~xsuchom2/papers/PomikalekSuchomel_SpiderlingEfficiency.pdf (dostop 20. 3. 2015).

Suhadolnik, Stane, 1968: Koncept novega slovarja slovenskega knjižnega jezika. *4. Seminar slovenskega jezika, literature in kulture. Predavanja iz jezika.* 1–11.

Suhadolnik, Stane, 1997: Šolar-Ramovšev načrt za slovar knjižnega jezika. *Slavistična revija* 45/3-4. 558–566.

Summers, Della, 1988, The Role of Dictionaries in Language Learning. Carter, Ron in Michael McCarthy (ur.): *Vocabulary and Language Teaching*. Longman. 111–125.

Svartvik, Jan, 1992: Lexis in English language corpora. Tommola, Hannu, Krista Varantola, Tarja Salmi-Tolonen in Jurgen Schopp (ur.): *EURALEX ,92 Proceedings I-II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland. Studia translatologica.* 17–31.

Šeruga Prek, Cvetka in Emica Antončič, 2003: *Slovenska zborna izreka. Priročnik z vajami za javne govorce*. Maribor: Aristej.

Ševčíková, Magda, Zdeněk Žabokrtský in Oldřich Krůza, 2007: Named Entities in Czech: Annotating Data and Developing NE Tagger. *Text, Speech and Dialogue. Lecture Notes in Computer Science* 4629. 188–195.

Šimková, Mária in Radovan Garabík, 2014: Slovenský národný korpus (2002–2012): východiská, ciele a výsledky pre výskum a prax. Gajdošová, Katarína in Adriána Žáková (ur.): *Jazykovedné štúdie XXXI: Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu)*. Bratislava: VEDA. 35–64.

Škiljan, Dubravko, 1999: *Javni jezik: k lingvistiki javne komunikacije*. Ljubljana: Studia Humanitatis.

Škofic, Jožica, 2009: Oznake za izgovorjavo v novem slovarju slovenskega jezika. Perdih, Andrej (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 135–143.

Šnajder, Jan, 2013: Models for predicting the inflectional paradigm of Croatian words. *Slovenščina 2.0* 1/2. 1–34. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_02.pdf (dostop 1. 8. 2015).

Šorli, Mojca (ur.), 2012a: *Dvojezična korpusna leksikografija. Slovenščina v kontrastu: novi obeti, novi izzivi*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Šorli, Mojca, 2012b: Pragmatični pomen in semantična prozodija v medjezični perspektivi: primer slovenščine in angleščine. Vintar, Špela (ur.): *Prevodi skozi korpusno prizmo*. Ljubljana: Znanstvena založba Filozofske fakultete. 180–205.

Šorli, Mojca, 2012c: Semantična prozodija v teoriji in praksi – korpusni pristop k proučevanju pragmatičnega pomena: primer slovenščine in angleščine. Šorli, Mojca (ur.): *Dvojezična korpusna leksikografija. Slovenščina v kontrastu: novi obeti, novi izzivi*. Ljubljana: Trojina, zavod za uporabno slovenistiko. 90–116.

Šorli, Mojca, 2013: Jezikovna in funkcijska zvrstnost v leksikalni podatkovni zbirki: kvalifikator ali pomenski opis? Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete UL. 435–441.

Šorli, Mojca, 2014a: *Pragmatični pomen v dvojezičnem slovaropisju*. Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Šorli, Mojca, 2014b: Sodobni sporazumevalni slovar slovenskega jezika: izhodišča, viri, izvedba. Grahek, Irena in Simona Bergoč (ur.): *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo*. Ljubljana: Ministrstvo za kulturo RS. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/23-_Mojca_Sorli_-Posvet_PSSJ-MZK-17-03-14-koncna.pdf (dostop 10. 8. 2015).

Štajner, Tadej, Tomaž Erjavec in Simon Krek, 2013: Razpoznavanje imenskih entitet v slovenskem besedilu. *Slovenščina 2.0* 1/2. 58–81. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_04.pdf (dostop 15. 1. 2015).

Štebe, Janez, Sonja Bezjak in Sonja Lužar, 2013: *Odprti podatki: načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji*. Ljubljana: Založba Fakultete za družbene vede.

Štícha, František in sodelavci, 2013: *Akademická gramatika spisovné češtiny.* Praha: Academia.

Šuštaršič, Rastislav, Smiljana Komar in Bojan Petek, 1995: Slovene. *Journal of the International Phonetic Association* 25. 86–90.

Šuštaršič, Rastislav, Smiljana Komar in Bojan Petek, 1999: Slovene. *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press. 135–139.

Švedova, Natalija Jul›evna in sodelavci, 1970: *Grammatika sovremmenogo russkogo literaturnogo jazyka.* Moskva: Izdatel'stvo Nauka.

Tarp, Sven, 2009: Reflections on lexicographical user research. *Lexikos* 19/1. 275–296.

Tavčar, Aleš, Darja Fišer in Tomaž Erjavec, 2012: sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Inštitut Jožef Stefan. 197–202.

Taylor R., John, 2003: Polysemy's paradoxes. *Language Sciences* 25. 637–655.

TEI P5: Guidelines for Electronic Text Encoding and Interchange, 2013. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (dostop 15. 1. 2015).

Teubert, Wolfgang, 2005: Korpusno jezikoslovje in leksikografija. Gorjanc, Vojko in Simon Krek (ur.): Študije o korpusnem jezikoslovju. Ljubljana: Krtina. 103–136.

Thomas, George, 1991: *Linguistic purism*. London in New York: Longman.

Thomas, James, 2015: *Discovering English with Sketch Engine*. Verstile.

Tiberius, Carole in Simon Krek, 2014: *Workflow of Corpus-Based Lexicography*. Deliverable COST-ENeL-WG3 meeting July 2014. Bolzano/Bozen.

Tiberius, Carole in Tanneke Schoonheim, 2015: The Algemeen Nederlands Woordenboek (ANW) and its Lexicographical Process. Hildenbrandt, Vera (ur.): *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie".* Mannheim: Institut für Deutsche Sprache.

Tiberius, Carole, Kris Heylen in Simon Krek, 2015: *Automatic Knowledge Acquisition for Lexicography*. Deliverable COST-ENeL-WG3 meeting August 2015. Herstmoceux, UK.

Tivadar, Gorazd in Hotimir Tivadar, 2015: Problematika snemanja in zapisovanja govorjenih besedil na slovenskih sodiščih (na primeru sojenja na okrajnem sodišču v severovzhodni Sloveniji). Tivadar, Hotimir (ur.): *Država in narod v slovenskem jeziku, literaturi in kulturi. 51. seminar slovenskega jezika, literature in culture*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 43–50.

Tivadar, Hotimir in Peter Jurgec, 2003: Podoba govorjenega slovenskega knjižnega jezika v Slovenskem pravopisu 2001. *Slavistična revija* 51/2. 203–220.

Tivadar, Hotimir, 2004a: Podoba in funkcija govorjenega knjižnega jezika glede na neknjižne zvrsti. Kržišnik, Erika (ur.): *Aktualizacija jezikovnozvrstne teorije na Slovenskem: členitev jezikovne resničnosti*. *Obdobja* 22. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. 437–452.

Tivadar, Hotimir, 2004b: Priprava, izvedba in pomen perceptivnih testov za fonetično-fonološke raziskave (na primeru analize fonoloških parov). *Jezik in slovstvo* 49/2. 17–36.

Tivadar, Hotimir, 2007: Vprašljivost nekaterih "večnih resnic" v govorjenem knjižnem jeziku – na primeru samoglasnikov. *Acta Universitatis Carolinae. Philologica. Phonetica Pragensia* 11. 59–74.

Tivadar, Hotimir, 2008: *Kakovost in trajanje samoglasnikov v govorjenem knjižnem jeziku*. Doktorska disertacija. Ljubljana: Filozofska fakulteta UL.

Tivadar, Hotimir, 2010: Normativni vidik slovenščine v 3. tisočletju – knjižna slovenščina med realnostjo in idealnostjo. *Slavistična revija* 58/1. 105–116.

Tomaszczyk, Jerzy, 1979: Dictionaries: users and uses. *Glottodidactica* 12. 103–119.

Tono, Yukio, 1984: *On the dictionary user's reference skills*. B. Ed. Dissertation. University of Tokyo.

Tono, Yukio, 2000: On the effects of different types of electronic dictionary interfaces on L2 learners' reference behaviour in productive/receptive tasks. Heid, Ulrich, Stefan Evert, Egbert Lehmann in Christian Rohrer (ur.): *Proceedings of the 9th EURALEX International Congress, EURALEX 2000*. Stuttgart: Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart. 855–862.

Tono, Yukio, 2001: *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension*. Berlin in New York: de Gruyter.

Tono, Yukio, 2011: Application of Eye-Tracking in EFL Learners. Dictionary Look-up Process Research. *International Journal of Lexicography* 24/1. 124–153.

Toporišič, Jože, 1969: O eno- in večnaglasnosti nekaterih besednih kategorij. *Jezik in slovstvo* 14. 51–59.

Toporišič, Jože, 1971a: Pravopis, pravorečje in oblikoslovje v Slovarju slovenskega knjižnega jezika I. *Slavistična revija* 19/1. 55–75.

Toporišič, Jože, 1971b: Pravopis, pravorečje in oblikoslovje v Slovarju slovenskega knjižnega jezika I. *Slavistična revija* 19/2. 222–229.

Toporišič, Jože, 1974a: K izrazju in tipologiji slovenske frazeologije. *Jezik in slovstvo* 19/8. 273–279.

Toporišič, Jože, 1974b: Kratko oblikoslovje slovenskega knjižnega jezika. Kmecl, Matjaž, Tine Logar in Jože Toporišič (ur.): *Slovenski jezik, literatura in kultura. Informativni zbornik.* Ljubljana: Filozofska fakulteta. 29–50.

Toporišič, Jože, 1974/75a: Besednovrstna vprašanja slovenskega knjižnega jezika. *Jezik in slovstvo* 20/2-3. 33–39.

Toporišič, Jože, 1974/75b: Esej o slovenskih besednih vrstah. *Jezik in slovstvo* 20/8. 295–305.

Toporišič, Jože, 1975: Formanti slovenskega knjižnega jezika. *Slavistična revija* 23. 153–196.

Toporišič, Jože, 1976 (1984, 2000, 2004): *Slovenska slovnica.* Maribor: Obzorja.

Toporišič, Jože, 1982: *Nova slovenska skladnja*. Ljubljana: Državna založba Slovenije.

Toporišič, Jože, 1988: Jezikoslovje s simpozija Obdobja 8. *Slavistična revija* 36/4. 437–449.

Toporišič, Jože, 1991: *Družbenost slovenskega jezika. Sociolingvistična razpravljanja*. Ljubljana: Državna založba Slovenije.

Toporišič, Jože, 1992: *Enciklopedija slovenskega je*zika. Ljubljana: Cankarjeva založba.

Toporišič, Jože, 2003: *Oblikoslovne razprave.* Ljubljana: Založba ZRC, ZRC SAZU.

Toporišič, Jože, 2007: Jezikoslovni slovnični teoremi Jožeta Toporišiča. Orel, Irena (ur.): *Razvoj slovenskega strokovnega jezika. Obdobja* 24. Ljubljana: Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik, Filozofska fakulteta UL. 401–413.

Toporišič, Jože, 2009: K posvetu o novem slovarju slovenskega jezika. *Slavistična revija* 57/4. 629–631.

Toporišič, Jože, 2011: *Intervjuji in polemike.* Ljubljana: ZRC SAZU.

Trap-Jensen, Lars, 2004: Spoken language in dictionaries: Does it really matter? Williams, Geoffrey in Sandra Vessier (ur.): *Proceedings ofthe Eleventh EURALEX International Congress. EURALEX 2004*. Lorient: Université de Bretagne-Sud. 311–318.

Trap-Jensen, Lars, Henrik Lorentzen in Nicolai Sørensen, 2014: An odd couple – Corpus frequency and look-up frequency: what relationship? *Slovenščina 2.0* 2/2. 94–113. http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_07.pdf (dostop 22. 6. 2015).

Trésor de la langue française, 1971–1994. Nancy: Inalf/Atilf.

Učni načrt. Program Osnovna šola. Slovenščina, 2011: http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/podrocje/os/prenovljeni_UN/UN_slovenscina_OS.pdf (dostop 22. 6. 2015).

Urdang, Laurence, 1984: A lexicographer's adventures in computing. *Dictionaries: Journal of the Dictionary Society of North America* 6.1. 150–165.

Uršič, Marko in Olga Markič, 1997: *Osnove logike*. Ljubljana: Filozofska fakulteta.

Varantola, Krista, 2002: Use and Usability of Dictionaries: Common Sense and Context Sensibility? Corréard, Marie-Hélène (ur.): *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Grenoble: EURALEX. 30–44.

Venhuizen, Noortje J., Valerio Basile, Kilian Evang in Johan Bos, 2013: Gamification for word sense labeling. Erk, Katrin in Alexander Koller (ur.): *Proceedings of the 10th International Conference on Computational Semantics. IWCS 2013*. Potsdam, Nemčija. 397–403.

Verdonik, Darinka, Matej Rojc, Zdravko Kačič in Bogomir Horvat, 2002: Zasnova in izgradnja oblikoslovnega in glasovnega slovarja za slovenski knjižni jezik. Tomaž Erjavec in Jerneja Gros (ur.): *Zbornik konference Jezikovne tehnologije 2002*. Ljubljana: Institut Jožef Stefan. 44–48.

Verdonik, Darinka in Matej Rojc, 2004: Jezikovni viri projekta LC-STAR. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Jezikovne tehnologije: zbornik 7. mednarodne multi-konference Informacijska družba IS 2004.* Ljubljana: Institut Jožef Stefan. 24–47.

Verdonik, Darinka, Matej Rojc in Zdravko Kačič, 2004: Creating Slovenian language resources for development of speech-to-speech translation components. *Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC'04*. Lisbon, Portugal. 1399–1402.

Verdonik, Darinka, 2007: *Jezikovni elementi spontanosti v govoru: Diskurzni označevalci in popravljanja*. Maribor: Slavistično društvo.

Verdonik, Darinka in Ana Zwitter Vitez, 2011: *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Verdonik, Darinka in Zdravko Kačič, 2012: Pragmatic functions of Christian expressions in spoken discourse. *Linguistica* 52. 267–281.

Verdonik, Darinka, Iztok Kosem, Ana Zwitter Vitez, Simon Krek in Marko Stabej, 2013: Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation* 47/4. 1031–1048.

Verdonik, Darinka, 2015a: Govorjeni proti pisnemu ali katera leksika je »tipično govorjena«. Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 392–405.

Verdonik, Darinka, 2015b: Jezikovnoteoretska načela v korpusnem jezikoslovju. *Slovenščina 2.0* 3/1. 1–27. http://www.trojina.org/slovenscina2.0/arhiv/2015/1/Slo2.0_2015_1_01.pdf (dostop 28. 7. 2015).

Verlinde, Serge in Jean Binon, 2010: Monitoring dictionary use in the electronic age. Dykstra, Anne in Tanneke Schoonheim (ur.): *Proceedings of the XIV EURALEX International Congress*. Leeuwarden/Ljouwert: Fryske Akademy-Afûk. 1144–1151.

Verovnik, Tina, 2013: Povej mi, kako zvenijo dnevnoinformativne oddaje, in povem ti, za kateri radijski program gre. Žele, Andreja (ur.): *Družbena funkcijskost jezika: vidiki, merila, opredelitve. Obdobja* 32. Ljubljana: Znanstvena založba Filozofske fakultete. 241–246.

Vicknair, Chad, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen in Dawn Wilkins, 2010: A Comparison of a Graph Database and a Relational Database: A Data Provenance Perspective. *Proceedings of the 48th Annual Southeast Regional Conference* 42. New York: ACM.

Vičič, Jernej, 2012: *Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike*. Doktorska disertacija. Ljubljana: Fakulteta za racunalništvo in informatiko UL.

Vidovič Muha, Ada, 1972: Oris dveh osnovnih pojavnih oblik sistema knjižnega jezika: (ob Slovarju slovenskega knjižnega jezika I). *Jezik in slovstvo* 17/6. 178–186.

Vidovič Muha, Ada, 1978: Merila pomenske delitve nezaimenske pridevniške besede. *Slavistična revija* 26/3. 253–276.

Vidovič Muha, Ada, 1992: Normativnost v slovar ujete slovenske besede. *Naši razgledi* 41/1 (10. jan. 1992). 10–11.

Vidovič Muha, Ada, 1996: Določnost kot besedilna prvina v slovničnem opisu slovenskega jezika (Ob Kopitarjevi slovnici). Toporišič, Jože (ur.): *Kopitarjev zbornik. Obdobja* 15. Ljubljana: Filozofska fakulteta. 115–130.

Vidovič Muha, Ada, 1997: Razmerja med leksemi in homonimija. Bálint, Júlia: *Slovar slovenskih homonimov: Na podlagi gesel Slovarja slovenskega knjižnega jezika*. Ljubljana: Znanstveni inštitut Filozofske fakultete. 7–16.

Vidovič Muha, Ada, 2000: *Slovensko leksikalno pomenoslovje. Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete UL.

Vidovič Muha, Ada, 2003: Sodobni položaj nacionalnih jezikov v luči jezikovne politike. Vidovič Muha, Ada (ur.): *Slovenski knjižni jezik – aktualna vprašanja in zgodovinske izkušnje. Obdobja* 20. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. 5–25.

Vidovič Muha, Ada, 2007: Izrazno-pomenska tipologija poimenovanj. *Slavistična revija* 55/1-2. 399–406.

Vidovič Muha, Ada, 2009: Poskus določitve meril slovarskega pomena. Perdih, Andrej (ur.): *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU. 27–36.

Vidovič Muha, Ada, 2013a: *Moč in nemoč knjižnega jezika*. Ljubljana: Znanstvena založba FF UL.

Vidovič Muha, Ada, 2013b: *Slovensko leksikalno pomenoslovje.* Ljubljana: Znanstvena založba Filozofske falkultete UL.

Vikør, Lars S., 2009: Lexicography and language planning in Scandinavia and the Netherlands. Nielsen, Sandro in Sven Tarp (ur.): *Lexicography in the 21st century. In honour of Henning Bergenholtz.* Amsterdam in Philadelphia: John Benjamins. 123−143.

Vintar, Špela in Tomaž Erjavec, 2008: iKorpus in luščenje izrazja za Islovar. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 65–69.

Vintar, Špela, 2009: Samodejno luščenje terminologije – izkušnje in perspektive. Ledinek, Nina, Mojca Žagar Karer in Marjeta Humar (ur.): *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU. 345–356.

Vintar, Špela in Darja Fišer, 2009: Adding Multi-Word Expressions to sloWNet. Erjavec, Tomaž (ur.): *Mondilex Fifth Open Workshop: Research Infrastructure for Digital Lexicography: Proceedings of the 12th International Multiconference Information Society*. Ljubljana: Institut Jožef Stefan. 56–63.

Vintar, Špela, 2010: Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16/2. 141–158.

Vintar, Špela, 2015a: Analiza iskalnih poizvedb na portalu Termania. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 434–442.

Vintar, Špela, 2015b: Specializirana leksika v splošnem slovarju. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 408–422.

Vintar, Špela in Nataša Logar, 2015: Luščenje specializiranih izrazov za splošni slovar. Gorjanc, Vojko, Polona Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 424–433.

Vrbinc, Marjeta in Alenka Vrbinc, 2014: Differences in the Inclusion and Treatment of Terminology in OALD3, OALD4 and OALD8. *Lexikos* 23/1. 440–455.

Web Content Accessibility Guidelines 2.0. http://www.w3.org/TR/WCAG20/ (dostop 12. 6. 2015).

Weinrich, Uriel, 1970: La définition lexicographique dans la sémantique descriptive. *Langages* 19. 69–86.

Wiegand, Herbert E., 1987: Zur handlungstheoretischen Grundlegung der Wörterbuchbenutzungsforschung. *Lexicographica. International Annual for Lexicography* 3. 178–227.

Wiegand, Herbert E., 1989: Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch. Hausmann, Franz J., Oskar Reichmann, Herbert E. Wiegand in Ladislav Zgusta (ur.): *Wörterbucher. Ein internationales Handbuch zur Lexikographie*. 1. knjiga. Berlin in New York: de Gruyter. 530–588.

Wiegand, Herbert E., 1992: Elements of a Theory towards a so-called Lexicographic Definition. *Lexicographica* 8. 175–289.

Williams, John, 1996: Enough Said: The Problems of Obscurity and Cultural Reference in Learner's Dictionary Examples. Martin Gellerstam (ur.): *Euralex ,96 Proceedings I-II*. Gothenburg: Gothenburg University. 497–505.

Winkler, Birgit, 2001: English learners' dictionaries on CD-ROM as reference and language learning tools. *ReCALL* 13/2. 191–205.

Wolfer, Sascha, Alexander Koplenig, Peter Meyer in Carolin Müller-Spitzer, 2014: Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses. Abel, Andrea, Chiara Vettori in Natascia Ralli (ur.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism. 281–290.

Wood, Peter T., 2012: Query Languages for Graph Databases. *ACM SIGMOD Record* 41/1. 50–60.

Wooldridge, Terence R., 1992: Structures du Corpus et de la Base Estienne-Nicot (1531-1628). *CCH Working Papers* 2. 21–32.

Wright, Jon, 1998: *Dictionaries*. Oxford: Oxford University Press.

Yeroshina Pobirk, Olga, Petra Zaranšek in Simon Šuster, 2009: Besedne skice za slovenščino. Kritični pogled. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 413–422.

Zaidan, Omar F. in Chris Callison-Burch, 2011: Crowdsourcing Translation: Professional Quality from Non-Professionals. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, ZDA. 1220–1229.

Zakon o avtorskih in sorodnih pravicah (ZASP), 2007. *Uradni list RS* 16. https://www.uradni-list.si/1/content?id=78529 (dostop 1. 8. 2015).

Zgusta, Ladislav, 1971: *Manual of Lexicography*. The Hague.

Žagar, Igor Ž. in Mojca Schlamberger Brezar, 2009: *Argumentacija v jeziku.* Ljubljana: Pedagoški inštitut.

Žaucer, Rok in Franc Marušič, 2009: *Jezikovno svetovanje, praksa in ideali*. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja* 28. Ljubljana: Znanstvena založba Filozofske fakultete UL. 449–456.

Žele, Andreja, 2001: *Vezljivost v slovenskem jeziku.* Ljubljana: Založba ZRC, ZRC SAZU.

Žele, Andreja, 2003a: *Glagolska vezljivost: iz teorije v slovar.* Ljubljana: Založba ZRC, ZRC SAZU.

Žele, Andreja, 2003b: Slovenska skladnja z vidika skladenjskih teorij. *Slavistična revija* 51/pos. št. 141–163.

Žele, Andreja, 2012a: Konektorji v slovenščini. *Zbornik Matice srpske za slavistiku* 62. 59–69.

Žele, Andreja, 2012b: *Pomensko-skladenjske lastnosti slovenskega glagola.* Ljubljana: Založba ZRC, ZRC SAZU.

Žele, Andreja, 2014a: Členki tudi kot vnašalniki novih prostorskih razmerij v obstoječe sporočilo. *Slavistična revija* 62/3. 321–330.

Žele, Andreja, 2014b: *Slovar slovenskih členkov*. Ljubljana: Založba ZRC, ZRC SAZU.

Žele, Andreja, 2015: Slovar členkov kot živa vez med besedilom in slovarjem. *Slavia Centralis* 8/1. 16–21.

Železnikar, Jaka, 1998: FIDA – pogoste napake pri vnosu in obdelavi besedil ter njihovo odpravljanje. *Uporabno jezikoslovje* 6. 107–111.

Žganec Gros, Jerneja, Varja Cvetko-Orešnik in Primož Jakopin, 2006: SI-PRON: a comprehensive pronunciation lexicon for Slovenian. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006*. Ljubljana: Institut „Jožef Stefan". 44–49.

Žgank, Andrej, Darinka Verdonik in Zdravko Kačič, 2008: Slovenska baza BNSI broadcast news za razpoznavanje tekočega govora. *Elektrotehniški vestnik* 75/3. 85–90.

Żmigrodzki, Piotr, 2010: Polish Academy of Sciences Great Dictionary of Polish History, precence, prospects. *Studies in Polish Linguistics* 6. 7–25.

Żmigrodzki, Piotr, 2014a: Polish Academy of Sciences Great Dictionary of Polish – Lexicographical Workflow and Summary of the Project. *Workflow of Corpus-based Lexicography*, COST ENeL WG3 meeting, Bolzano, 19 julij. http://www.elexicography.eu/wp-content/uploads/2014/07/Zmigrodzki_2014_COST_Bolzano.pdf (dostop 6. 7. 2015).

Żmigrodzki, Piotr, 2014b: Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. *Slovenščina 2.0* 2/2. 37–52. http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_04.pdf (dostop 10. 8. 2015).

Spletni jezikovni viri in orodja (dostop 1. 8. 2015)

ANW: Algemeen Nederlands Woordenboek. http://anw.inl.nl/
Collins. http://www.collinsdictionary.com/
Daele: Diccionario de aprendizaje de español como lengua extranjera. http://www.daele.eu/
DDO: Den Danske Ordbog. http://ordnet.dk/ddo
Diccionario de la lengua Española. http://lema.rae.es/drae
DWDS: Das Digitale Wörterbuch der deutschen Sprache. http://www.dwds.de/
Eesti keele seletav sõnaraamat. http://en.eki.ee/dict/ekss
elexiko: Online-Wörterbuch zur deutschen Gegenwartssprache. http://www.owid.de/wb/elexiko/start.html
Evroterm: večjezična terminološka zbirka. http://www.evroterm.gov.si/
Folkets lexikon. http://folkets-lexikon.csc.kth.se/
Groningen Meaning Bank. http://gmb.let.rug.nl/
Hrvatski enciklopedijski rječnik. http://hjp.novi-liber.hr
Interactive Language Toolbox. https://ilt.kuleuven.be/inlato/
Islovar. http://www.islovar.org
Jezikovni svetovalni servis. http://www.rtvslo.si/podcasts/svetovalni_serivs.xml
Knjižnica prosto dostopnih slovarjev. http://www.evroterm.gov.si/slovar/index.html
Kolaborativni spletni portal Wiktionary. https://www.wiktionary.org/
Korpus govorjene slovenščine Gos. http://www.slovenscina.eu/korpusi/gos
Korpus slovenskega jezika FidaPLUS. http://www.fidaplus.net
Korpus slovenskega jezika Gigafida. http://www.slovenscina.eu/korpusi/gigafida
Korpus starejše slovenščine IMP. http://nl.ijs.si/imp/
Krvina, Domen: *Sprotni slovar slovenskega jezika*. Različica 1.0. http://fran.si/132/sssj-sprotni-slovar-slovenskega-jezika
Leksikalna baza za slovenščino. http://www.slovenscina.eu/spletni-slovar/leksikalna-baza
Lexical Markup Framework. http://en.wikipedia.org/wiki/Lexical_Markup_Framework
Longman Dictionary of Contemporary English. http://www.ldoceonline.com/
Macmillan English Dictionary. http://www.macmillandictionary.com/
Oxford dictionaries. http://www.oxforddictionaries.com/
Pedagoški slovnični portal. http://www.slovenscina.eu/portali/pedagoski-slovnicni-portal
Pleteršnik, Maks, 1894–1895: *Slovensko-nemški slovar*. Inštitut Frana Ramovša, ZRC SAZU. http://www.fran.si/136/maks-pletersnik-slovensko-nemski-slovar
Predstavitvena verzija pregibnika Amebis Besana. http://besana.amebis.si/pregibanje/
Presisov večjezični slovar, 2013: Kamnik: Amebis d. o. o. http://www.termania.net/slovarji/70/presisov-vecjezicni-slovar
Razvezani jezik. Prosti slovar žive slovenščine. http://razvezanijezik.org/
Slovenski oblikoslovni leksikon Sloleks. http://www.slovenscina.eu/sloleks/
Slovenski semantični leksikon sloWNet. http://nl.ijs.si/slownet
Slovník spisovného jazyka českého. http://ssjc.ujc.cas.cz
SNB: Slovar novejšega besedja slovenskega jezika. http://www.fran.si/131/snb-slovar-novejsega-besedja
SP: Slovenski pravopis, 2001. http://bos.zrc-sazu.si/sp2001.html

SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/ (dostop 12. 6. 2015).
Spletni portal Fran: slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU. http://www.fran.si
Spletni slovar slovar slovenskega jezika/Leksikalna baza za slovenščino. (http://www.slovenscina.eu/spletni-slovar)
SSKJ: Slovar slovenskega knjižnega jezika. http://www.fran.si/130/sskj-slovar-slovenskega--knjiznega-jezika
SSKJ2: Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja. http://www.sskj2.si/
Statistics used in the Sketch Engine, 2013. Lexical Computing Ltd. http://www.sketchengine.co.uk
Šolar – korpus šolskih pisnih besedil. http://www.slovenscina.eu/korpusi/solar
ŠUSS: Odgovori na jezikovna vprašanja. http://www2.arnes.si/~lmarus/suss/index.html
Termania. http://www.termania.net
Terminologišče. http://isjfr.zrc-sazu.si/en/terminologisce#v
Text Encoding Initiative. http://www.tei-c.org
The Merriam-Webster Online Dictionary. http://www.merriam-webster.com
TheFreeDictionary. http://www.thefreedictionary.com/
Urban Dictionary. http://www.urbandictionary.com/
Vocabulary.com. http://www.vocabulary.com/
Wielki słownik języka polskiego PAN. http://www.wsjp.pl
Wikislovar. http://sl.wiktionary.org/
Wordsmyth.net. http://www.wordsmyth.net

Statistični podatki SURS (dostop 12. 6. 2015)

Podatki SURS (2015a). *Prebivalstvo, staro 15 ali več let, po izobrazbi, starosti in spolu, Slovenija, letno.* http://pxweb.stat.si/pxweb/Dialog/varval.asp?ma=05G2002S&ti=&path=../Database/Dem_soc/05_prebivalstvo/20_soc_ekon_preb/01_05G20_izobrazba/&lang=2

Podatki SURS (2015b). *Dijaki po starosti, letnikih, spolu in vrsti izobraževanja, Slovenija, letno.* http://pxweb.stat.si/pxweb/Dialog/varval.asp?ma=0953201S&ti=&path=../Database/Dem_soc/09_izobrazevanje/07_srednjesol_izobraz/01_09532_zac_sol_leta/&lang=2

Podatki SURS (2015c). *Vpis učencev v osnovnošolsko izobraževanje po vrsti šole, vrsti programa in spolu, Slovenija, začetek šolskega leta, letno.* http://pxweb.stat.si/pxweb/Dialog/varval.asp?ma=0952701S&ti=&path=../Database/Dem_soc/09_izobrazevanje/04_osnovnosol_izobraz/01_09527_zac_sol_leta/&lang=2

Podatki SURS (2015d). *Obseg opremljenosti gospodinjstev z informacijsko-komunikacijsko opremo po tipu gospodinjstva, Slovenija, večletno*. http://pxweb.stat.si/pxweb/Dialog/varval.asp?ma=2973901S&ti=&path=../Database/Ekonomsko/23_29_informacijska_druzba/10_IKT_gospodinjstva/02_29739_opremljenost_IKT/&lang=2

Podatki SURS (2015e). *Vzroki za nedostop do interneta po tipu gospodinjstva, Slovenija, letno.* http://pxweb.stat.si/pxweb/Dialog/varval.asp?ma=2974003S&ti=&path=../Database/Ekonomsko/23_29_informacijska_druzba/10_IKT_gospodinjstva/04_29740_dostop_internet/&lang=2

Podatki SURS (2015f). *Pogostost in kraj uporabe interneta pri posameznikih po starostnih razredih in spolu, Slovenija, letno.* http://pxweb.stat.si/pxweb/Dialog/varval.asp?ma=2974201S&ti=&path=../Database/Ekonomsko/23_29_informacijska_druzba/11_IKT_posamezniki/04_29742_uporaba_inter/&lang=2

Poročilo SURS (2015a). *1. januarja 2015 Slovenija s 2.062.874 prebivalci, 5 % tujih državljanov.* http://www.stat.si/StatWeb/prikazi-novico?id=5148

Poročilo SURS (2015b). *Uporabniki slovenskih mobilnih operaterjev so v letu 2014 poslali skoraj 2,4 milijarde SMS-sporočil.* http://www.stat.si/StatWeb/prikazi-novico?id=5083

Poročilo SURS (2015c*). Z naraščanjem uporabe IKT se razvijajo nove storitve in nastajajo nove potrebe.* http://www.stat.si/StatWeb/prikazi-novico?id=5184

Drugi spletni viri (dostop 1. 8. 2015)

Amazon Mechanical Turk. https://www.mturk.com/
Center za slovenščino kot drugi/tuji jezik. http://www.centerslo.net/
Clickworker. http://www.clickworker.com/en/
Creative Commons. https://creativecommons.org/
Crowdcrafting. http://crowdcrafting.org/
CrowdFlower. http://www.crowdflower.com/
Delo. http://www.delo.si/
Dnevnik. https://www.dnevnik.si/
Društvo ljubiteljskih pravopisarjev in slovničarjev. https://www.facebook.com/groups/191388157545784/
Družina. http://www.druzina.si/
Duolingo. https://www.duolingo.com/
FoldIt. https://fold.it/portal/
Igra besed. http://www.igra-besed.si/
Phrase Detectives. http://anawiki.essex.ac.uk/phrasedetectives/
Phylo. http://phylo.cs.mcgill.ca/
Prevajalci, na pomoč! https://www.facebook.com/groups/help.prevajalci/?fref=ts
PyBossa. http://pybossa.com/
RTV Slovenija. http://www.rtvslo.si/
sloWCrowd. http://nl.ijs.si/slowcrowd/
Wordrobe. http://wordrobe.housing.rug.nl/
Za vsaj približno pravilno rabo slovenščine. https://www.facebook.com/groups/398216690214010/

Seznam avtorjev

Ime in priimek	Elektronska pošta	Afiliacija
Špela Arhar Holdt	spela.arhar@trojina.si	Trojina, zavod za uporabno slovenistiko in Univerza v Ljubljani, Filozofska fakulteta, Oddelek za slovenistiko
Tatjana Balažic Bulc	tatjana.balazic-bulc@guest.arnes.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za slavistiko
Ciril Bohak	ciril.bohak@fri.uni-lj.si	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Jaka Čibej	jaka.cibej@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Kaja Dobrovoljc	kaja.dobrovoljc@trojina.si	Trojina, zavod za uporabno slovenistiko
Tomaž Erjavec	tomaz.erjavec@ijs.si	Institut »Jožef Stefan«, Odsek za tehnologije znanja
Ina Ferbežar	ina.ferbezar@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik
Darja Fišer	darja.fiser@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Luka Fürst	luka.fuerst@fri.uni-lj.si	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Polona Gantar	apolonija.gantar@guest.arnes.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Vojko Gorjanc	vojko.gorjanc@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Robert Grošelj	robert.groselj@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Peter Holozan	peter.holozan@amebis.si	Amebis, d. o. o.
Peter Jurgec	peter.jurgec@utoronto.ca	Univerza v Torontu, Oddelek za jezikoslovje in Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Monika Kalin Golob	monika.kalin-golob@fdv.uni-lj.si	Univerza v Ljubljani, Fakulteta za družbene vede
Bojan Klemenc	bojan.klemenc@fri.uni-lj.si	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Ime in priimek	Elektronska pošta	Afiliacija
Iztok Kosem	iztok.kosem@trojina.si	Trojina, zavod za uporabno slovenistiko in Center za jezikovne vire in tehnologije Univerze v Ljubljani
Simon Krek	simon.krek@guest.arnes.si	Institut »Jožef Stefan«, Laboratorij za umetno inteligenco in Center za jezikovne vire in tehnologije Univerze v Ljubljani
Nikola Ljubešić	nikola.ljubesic@ffzg.hr	Univerza v Zagrebu, Filozofska fakulteta, Odsek za informacijske in komunikacijske znanosti in Institut »Jožef Stefan«, Odsek za tehnologije znanja
Nataša Logar	natasa.logar@fdv.uni-lj.si	Univerza v Ljubljani, Fakulteta za družbene vede
Vesna Mikolič	vesna.mikolic@fhs.upr.si	Univerza na Primorskem, Fakulteta za humanistične študije, Znanstveno-raziskovalno središče
Gregor Perko	gregor.perko@guest.arnes.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za romanistiko
Nataša Pirih Svetina	natasa.pirih-svetina@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik
Agnes Pisanski Peterlin	neza.pisanski@guest.arnes.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Damjan Popič	damjan.popic@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Marko Robnik-Šikonja	marko.robniksikonja@fri.uni-lj.si	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Tadeja Rozman	tadeja.rozman@trojina.si	Trojina, zavod za uporabno slovenistiko in Univerza v Ljubljani, Filozofska fakulteta, Oddelek za slovenistiko
Marko Stabej	marko.stabej@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za slovenistiko
Mojca Šorli	mojca.sorli@siol.net	samostojna raziskovalka, leksikografinja in prevajalka
Darinka Verdonik	darinka.verdonik@um.si	Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Ime in priimek	Elektronska pošta	Afiliacija
Špela Vintar	spela.vintar@ff.uni-lj.si	Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo
Ana Zwitter Vitez	ana.zwitter@fhs.upr.si	Univerza na Primorskem, Fakulteta za humanistične študije in Univerza v Ljubljani, Filozofska fakulteta, Oddelek za prevajalstvo

Imensko kazalo

Iz recenzij

Monografija *Slovar sodobne slovenščine: problemi in rešitve* je impresiven interdisciplinarni raziskovalni dosežek, tematsko zelo izčrpen in koherenten prispevek tako k sodobni leksikografski teoriji kot tudi nadvse relevanten napotek za dinamično in interaktivno leksikografsko prakso. Gre za rezultat skupnega prizadevanja urednikov k osmišljanju koncepta novega enojezičnega slovarja sodobnega slovenskega jezika, in sicer z namenom, da se predvidijo in analizirajo ključni problem pri pripravi slovarja, pri njihovem reševanju pa se odgovarja tudi na izzive današnjih komunikacijskih potreb in tehnoloških možnosti.

Kompleksni problemi so obdelani v dvaintridesetih študijah, ki obsegajo širok tematski razpon, od definiranja sociolingvističnega konteksta in jezikovnopolitičnih značilnosti vloge slovarja kot pripomočka za jezikovno standardizacijo ter oblikovanja standardnojezikovne kulture do analiz dejanskih uporabnikov. Naloga načrtovanega slovarja je, da vsem skupinam uporabnikov poda jasen in zanesljiv leksikalno-slovnični opis slovenskega jezika, utemeljen na podatkih iz reprezentativnih jezikovnih korpusov, hkrati pa je njegov namen tudi, da s svojim potencialom postane izhodišče za razvoj jezikovnih tehnologij. Monografija na inovativen način podaja znanstveno razpravo o vrsti leksikografsko relevantnih tem: od normativnih vprašanj do predstavitve fontetičnih, oblikoskladenjskih in stilističnih informacij, od vloge zgledov v slovarju do slovarskih definicij. Posebno vrednost prinašajo poglavja o kriterijih za vključevanje strokovne leksike, o jezikovnotehnoloških postopkih in jezikovnotehnoloških orodjih, prav tako pa tudi razprave o potencialu moči množic pri leksikografskem delu.

Monografija je dragocen prispevek k vrsti jezikoslovnih področij, tudi k razpravi o aktualnih vprašanjih slovenske jezikovne norme, tako da ni namenjena zgolj jezikoslovcem, prevajalcem in študentom, ampak tudi širši strokovni javnosti.

Maja Bratanić
Inštitut za hrvaški jezik in jezikoslovje, Zagreb

Iz hrvaščine prevedel Vojko Gorjanc.

Glede monografije *Slovar sodobne slovenščine: problemi in rešitve* izrekam pohvalo mnogostranskemu pregledu različnih vidikov sestavljanja predlaganega slovarja. Avtorji, vešči jezikoslovno usmerjenega računalništva in seznanjeni z dogajanjem v evropskem in ameriškem korpusnem jezikoslovju ter korpusno zasnovani leksikografiji, ponujajo nove poglede in inovativne metode. Ideja ločevanja gradiva za slovar od samega slovarja ni nova, že oxfordski in drugi slovarski projekti so ustvarili »listkovne kartoteke« s citati; toda avtorji monografije so skrbno razmislili tako o tem, katere informacije (oblikoslovne, skladenjske, semantične, kolokacijske) bodo ohranjali v leksikalni bazi podatkov, ki naj služi različnim namenom, kot tudi o tem, kako naj bazo stalno posodabljajo z avtomatiziranim spremljanjem novosti v nenehno rastočem jeziku. Hkrati analizirajo tudi vse korake pri sestavljanju posameznega slovarskega članka in pri tem načrtujejo zadolžitve, za katere so potrebni profesionalni leksikografi, manjše naloge za mlajše specialiste, naloge, ki jih opravi računalnik, ter mikronaloge, ki jih pod nadzorom opravi izobražen nestrokovnjak v procesu množičenja. Če vsi izdelovalci slovarjev razmišljajo o idealiziranem bralcu, ta knjiga ponuja empirične študije uporabniških želja: mlajše generacije slovenskih in tujejezičnih uporabnikov ne potrebujejo tiskane knjige, celo ne predstavitve knjižnega formata na spletni strani, temveč za splet zasnovano vsebino, ki se lahko sproti prilagaja posameznim segmentom publike. Zaradi predlaganega *Slovarja sodobne slovenščine* se ne bom znebil obstoječega večzvezkovnega slovarja, s pomočjo katerega prebiram literaturo zadnjih dveh stoletij, upam pa, da bom kmalu brskal po njem na zaslonu, ko bom želel analizirati in interpretirati jezik našega časa.

Wayles Brown
Univerza Cornell, Ithaca, New York

Iz angleščine prevedel Simon Krek.

Ustvarjanje razlagalnega slovarja sodi med najzahtevnejše jezikoslovne projekte, ki si jih lahko predstavljamo. Če naj takšen priročnik ustreza sodobnim standardom, kot si prizadeva monografija *Slovar sodobne slovenščine: problemi in rešitve*, je treba začeti pri izdelavi jezikovnih virov, pripraviti koncept izdelave slovarja in na koncu poskrbeti še za tisto najtežje – sestaviti ekipo strokovnjakov, ki bo nekaj let slovar pripravljala. Pričujoča monografija je – ob ostalih dejavnostih (kot je organizacija danes že svetovno znanih srečanj v okviru konference eLex o digitalni leksikografiji) – še en dokaz, da je slovensko jezikoslovje odločno usmerjeno proti temu cilju.

Slovenske in češke jezikoslovne razmere so si v marsičem podobne, zato sta nam lahko v spodbudo obe smeri – ne češčina ne slovenščina nimata obsežne slovaropisne tradicije (primerljive npr. z angleško). Še v 90. letih je za oba jezika primanjkovalo podatkovnih zbirk za izdelavo slovarjev, kar je bilo preseženo z nastankom obsežnih korpusov. Temu ustreza tudi potreba po soočenju z dediščino preskriptivne tradicije, ki jo je treba v luči stvarnih jezikovnih podatkov odkloniti kot preseženo. Ta in mnoga druga vprašanja (tehnični vidiki slovarja, odnos uporabnikov, možnosti množičenja ipd.) so obravnavana v pričujoči knjigi, ki lahko spodbudi tudi obravnave drugih jezikov. Iskrene čestitke je treba zato pospremiti z željo, da bi ustvarjalci slovarja imeli dovolj moči do sklepne faze procesa, na koncu katerega se nahaja sodoben, elektronski, opisni razlagalni slovar sodobne slovenščine.

Václav Cvrček
Inštitut Češkega nacionalnega korpusa,
Filozofska fakulteta Karlove univerze v Pragi

Iz češčine prevedel Robert Grošelj.

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

In *Slovar sodobne slovenščine: problemi in rešitve* I applaud the manysided examination of ways to compile the proposed dictionary. Experienced in linguistically-oriented computation, informed about recent European and American corpus linguistics and corpus-based lexicography, the authors offer new viewpoints and innovative methods. Separating raw material for a dictionary from the dictionary itself is not new, since the Oxford and other dictionary projects create a 'slip file' of quotations; but the authors carefully consider both what information (morphological, syntactic, semantic, collocational) to keep in their database of lexical items letting it serve multiple purposes, and how to update it constantly through automated tracking of the evergrowing language. They, further, analyze the steps toward compiling each dictionary entry. They plan assignments requiring professional lexicographers, smaller tasks for younger specialists, those needing computer processing, and microtasks that educated laymen can do under due control in the 'crowdsourcing' process. All dictionary-makers have an ideal reader in mind, but the present book studies users' preferences empirically: younger generations of Slovenian and foreign users require not a printed book, not even a printed-book format presented on an internet page, but an internet-native presentation modifiable for each subset of the audience. The proposed Slovar sodobne slovenščine will not, for me, banish the existing multi-volume dictionary as I read literature of the last two centuries, but I hope soon to open it on my screen while analyzing and interpreting the language of our own years.

Wayles Browne
Cornell University, Ithaca, New York