Evalvacijska kategorizacija strojno izluščenih protipomenskih parov

Tina Mozetič,[*] Miha Sever,[*] Martin Justin,[*] Jasmina Pegan[‡]

[*] Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
tina.mozetic11@gmail.com, mihasever98@gmail.com, martin1123581321@gmail.com
[‡] Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
jp2634@student.uni-lj.si

Povzetek

Namen prispevka je oceniti relevantnost strojno pridobljenih protipomenskih parov za vključitev v razširjeni Slovar sopomenk sodobne slovenščine. Nekdanje strukturalistično pojmovanje protipomenskosti vedno bolj prehaja k sodobnejšemu, ki temelji na naprednih računalniških metodah, odprtosti, množičenju, relevantnosti in uporabnosti podatkov. V raziskavi smo pregledali 2852 strojno izluščenih parov protipomenk. Primeri, ki jih označevalci niso enoznačno uvrstili med protipomenske oziroma neprotipomenske, so razvrščeni v 21 kategorij. Za protipomenke vsake kategorije je opredeljeno, ali jih je smiselno vključiti v odzivni slovar. Strojni postopek se je izkazal za uspešnega, saj je v slovar mogoče vključiti 88 % izluščenih parov. Kategorije bodo v prihodnosti uporabne tudi za oblikovanje smernic ter razvoj nadaljnje metodologije strojnega luščenja protipomenk.

Evaluative Categorisation of Automatically Extracted Pairs of Antonyms

This paper aims to assess the relevance of extracted antonym pairs that are to be included in the expanded Thesaurus of Modern Slovene. The former structuralistic conception of antonymy is shifting to a more modern one that is based on advanced computational methods, openness, crowdsourcing, relevance, and data usability. In this study, we reviewed 2852 extracted pairs of antonyms. Examples that were not uniquely classified as antonyms or non-antonyms by the evaluators are grouped into 21 categories. For each category, it is determined whether they should be included in the responsive dictionary. The process proved to be successful, as 88% of the extracted pairs could be included in the dictionary. The categories will also be useful in the future for the creation of guidelines and the development of further methodologies for automatic extraction of antonyms.

1. Uvod

Slovar sopomenk sodobne slovenščine je s 105.473 iztočnicami in 368.117 sopomenkami »najobsežnejša prosto dostopna avtomatsko generirana zbirka sopomenk za slovenščino« (Sopomenke 1.0, 2022). Slovar deluje po principu odzivnega slovarja, ki je v prvem koraku pripravljen povsem strojno. Strojno pripravljeni podatki so objavljeni takoj, ko jezikoslovna evalvacija potrdi njihovo načelno ustreznost oz. relevantnost za skupnost, nato pa se slovar razvija naprej po korakih in v sodelovanju jezikoslovcev in širše zainteresirane javnosti (Arhar Holdt et al., 2018). Pri projektu Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL bomo sopomenkam dodali protipomenke, za katere je treba opraviti tovrstno jezikoslovno evalvacijo relevantnosti.

Cilj našega prispevka je tako oceniti relevantnost strojno pridobljenih protipomenskih parov za vključitev v razširjeni Slovar sopomenk sodobne slovenščine. Pri tem nas zanima predvsem, kateri del podatkov je (1) primeren za neposredno vključitev v slovar, (2) kateri za vključitev ni primeren in (3) kateri zahteva dodaten premislek. V prispevku se natančneje ukvarjamo s tretjo točko, pri čemer dokazujemo, da je »problematične« primere mogoče kategorizirati glede na vrsto problema in tako določiti, ali jih je (a) mogoče izboljšati strojno, ali (b) morda zahtevajo uredniško odločitev, (c) jih je mogoče izboljšati s pomenskim členjenjem gesla ali kvalifikatorji, (d) jih je mogoče izboljšati s pomočjo skupnosti oziroma (e) kljub določenemu problemu pustili v naboru slovarskega gradiva in računati na to, da bodo uporabniki sami presodili o njihovi uporabnosti.

Problemske kategorije, ki jih bomo tako oblikovali, bodo služile kot izhodišče za nadaljnje delo na projektu, ki obsega nadgradnjo metodologije luščenja, pripravo smernic za uredniško obravnavo protipomenk in vključitev protipomenk v Slovar sopomenk sodobne slovenščine. Ročno pregledani protipomenski pari bodo uporabljeni kot učna množica za nadaljnje luščenje protipomenk iz korpusa Gigafida 2.0 (Krek et al., 2020). Tudi pri oblikovanju smernic pa bo naša analiza prišla zelo prav, saj smo identificirali probleme, za katere bo treba v nadaljevanju podati tudi načelne uredniške rešitve.

V drugem razdelku prispevka tako najprej predstavimo jezikoslovne raziskave protipomenskosti in koncept odzivnega slovarja. V tretjem na kratko opišemo metodi pridobivanja in označevanja podatkov. V četrtem razdelku pa predstavimo rezultate označevanja in jih analiziramo. Najprej predstavimo odločitve označevalcev glede ustreznosti protipomenskih parov, nato pa natančneje predstavimo vsako od problemskih kategorij, v katere so bili v fazi označevanja uvrščeni »problematični« primeri. Pri vsaki kategoriji predstavimo tudi njeno pogostost in ocenimo, na kakšen način bi bilo identificirani problem mogoče reševati. V zaključnem delu povzamemo bistvene ugotovitve prispevka.

2. Pregled področja

Jezikoslovje smatra protipomenskost – poleg sopomenskosti – za temeljno medleksemsko pomensko razmerje (Stramljič Breznik, 2010; Humar, 2016; Vidovič Muha, 2005, 2021). V nasprotju s sopomenkami protipomenke nujno nastopajo binarno, tj. v parih, in so vedno del skupnega pojmovnega ali celo pomenskega polja (Vidovič Muha, 2021). V slovenskem izrazoslovju sta se

enakovredno ustalila izraza protipomenka in antonim oz. protipomenskost in antonimija, čeprav Slovenski pravopis 2001 prednost daje izrazu protipomenka (Humar, 2005).

Definiranje protipomenke je razmeroma enostavno. Protipomenka je po SSKJ (2014) »beseda z nasprotnim pomenom v odnosu do druge besede«, enako jo opredeljuje tudi Toporišič (2001). Marjeta Humar (2016) definicijo razširi na »poimenovanja pojmov z eno- ali večpomensko besedo ali besedno zvezo, [pri katerem] sta v protipomenskem razmerju pomenski sestavini pojmov (navadno po ena pri vsakem od dveh), izraženih z enopomenskima besedama, z enopomenskima besednima zvezama ali pa s posameznima pomenoma dveh večpomenskih besed ali zvez« (22).

V nasprotju z definiranjem pomenska tipološka razvrstitev protipomenk predstavlja veliko oviro; tovrstnih razvrstitev je namreč toliko, kolikor je znanstvenikov, ki so se z njimi ukvarjali. Problematike se zavedajo tudi jezikoslovci sami (gl. Humar, 2016), njihova glavna naloga pa bi bila določiti meje protipomenskosti (Gao in Zheng, 2014), ki se od enega do drugega znanstvenika močno razlikujejo.

Marjeta Humar (2016) med pionirske in najpomembnejše jezikoslovne raziskovalce protipomenskosti uvršča Lyonsa, Apresjana in Novikova. Lyons je določil tri vrste protipomenk, ki izhajajo iz ene od naštetih značilnosti: komplementarnost, protipomenskost in konverzija. Pri tem loči protipomenkost v ožjem in širšem smislu; v ožjega vključuje le polarno protipomenskost, ki je zanj najčistejša oblika antonimije. Apresjan je protipomenke razčlenil veliko temeljiteje, opozoril pa je tudi na kvaziprotipomenke, ki nimajo enako nasprotnih pomenov. Novikov je na drugi strani protipomenskost razdelil na kontrarno nasprotnost – kot najpogostejšo obliko, komplementarno in vektorsko nasprotnost. Med kvaziprotipomenke je uvrstil pomensko neenake, nesorazmerne, nesimetrične, stilistično raznorodne, časovno različne protipomenke, ki izražajo druga nasprotja.

V slovenskem prostoru se je najbolj uveljavila členitev po A. Vidovič Muha (2005, 2021), ki protipomenskost opredeljuje kot pomensko nasprotnost ali dopolnjevalno protislovnost; za izhodišče tipološke členitve jemlje vpliv protipomenk na aktantske vloge znotraj stavčne povedi. V okviru tega protipomenke deli na:

- zamenjavne oz. konverzivne,
- dopolnjevalne oz. komplementarne,
- skrajnostne oz. polarne, s podskupino stopnjevalnih oz. gradualnih in
- usmerjene oz. vektorske.

V grobem kategorizacija temelji torej bodisi na enakovrednih skupinah protipomenk bodisi na osi bolj protipomensko–manj protipomensko (ožji : širši smisel, prave protipomenke : kvaziprotipomenke, popolne : nepopolne, neostra : ostra nasprotnost, binarna : nebinarna nasprotnost, izražanje nasprotja : stilistično sredstvo) (Humar, 2016).

Strukturna delitev protipomenk je jasnejša. V slovenskem prostoru se je z njo z besedotvornega vidika največ ukvarjala Irena Stramljič Breznik (2010), ki protipomenke deli na istokorenske (tudi gramatične ali tvorbene) in raznokorenske (tudi leksikalne).

Slovenski, pa tudi sicer nekdanji jugoslovanski prostor protipomenskosti dolgo ni posvečal večje pozornosti (Humar, 2016), to izražajo tudi glavni slovenski jezikovni priročniki. SSKJ je s kvalifikatorjem ant. (antonim) opremil 87 leksemov, ki se uvrščajo med kakovostne (polarne, skrajnostne) protipomenke, medtem ko usmerjenih in dopolnjevalnih ne izkazuje (Humar, 2016). Toporišič (1976) v svoji slovnici antonimijo omenja bežno pri antonimnem pridevniku, protipomenskost krajše predstavi pozneje v četrti, prenovljeni izdaji leta 2001. Kljub temu da so protipomenke leksikografsko prepoznane kot pomemben dejavnik pri določanju pravilnih pomenov besed (Toporišič, 2001), protipomenskega slovarja v slovenskem prostoru še nimamo. Imamo pa dva slovarja sopomenk, in sicer Sinonimni slovar slovenskega jezika (SSSJ), ki ga je izdal ZRC SAZU, in spletni Slovar sopomenk sodobne slovenščine (SSSS), ki je nastal pod okriljem Centra za jezikovne vire in tehnologije. Pretekli leksikografski opis slovenskega jezika se je naslanjal na strukturalistično tradicijo SSKJ-ja, ki so ji sledili tudi dosedanji najvidnejši slovenski raziskovalci protipomenskosti (Jože Toporišič, Ada Vidovič Muha, Irena Stramljič Breznik, Marjeta Humar).

Družbene spremembe kot posledica digitalizacije in razvoja informacijsko-komunikacijske tehnologije so oblikovale potrebo po popolnoma drugačnem leksikografskem opisu slovenščine, na podlagi katerega bi lahko gradili nove jezikovne vire in tehnologije. Leksikografija se namreč v sedanjem času zaradi vstopa interneta spopada z vse hitrejšimi jezikovnimi spremembami. Na eni strani je soočena z vprašanjem, kako v spremenjenih razmerah predstaviti slovarske vsebine jezikovnim uporabnikom, na drugi strani pa z novimi jezikovnimi praksami, ki jih vse težje sproti zajema in popisuje (Gantar et al., 2016). Sodobni jezikovni uporabniki vse bolj zahtevajo takojšnji dostop do slovarskih vsebin sodobnega jezika, zato moramo leksikografske analize izvajati vse hitreje, a enako kvalitetno (Gantar et al., 2016). Iz tradicionalnega leksikografskega modela prehajamo v sodobnejši, pri katerem slovarske vsebine temeljijo na naprednih računalniških metodah, odprtosti, množičenju, relevantnosti in uporabnosti podatkov.

Tako je na eni strani povsem ročni pristop luščenja podatkov zamenjal polavtomatični, ki ni le časovno in finančno manj potraten, ampak hkrati zagotavlja dodatne potencialno koristne podatke za presojo o vključevanju leksemov v slovar. Pri tem se vloga leksikografa ne spreminja, saj še vedno ostaja odločevalec na vseh ravneh odločanja o slovarskem vključevanju leksemov, spreminja pa se način pridobivanja in predstavitve leksemskega podatka (Gantar et al., 2016). Podoben princip luščenja je bil uporabljen pri pripravi SSSS-ja. Leksemska razmerja navadno luščimo iz baze več virov, SSSS tako temelji na luščenju podatkov iz korpusa Gigafida in *Velikega angleško-slovenskega slovarja OXFORD - DZS* (Arhar Holdt et al., 2018). V tujini so pri pripravi korpusnih protipomenskih slovarjev prešli že na avtomatično luščenje (Wang et al., 2010; Lobanova et al., 2010; Aldhubayi in Alyahya, 2014).

Na drugi strani pa SSSS deluje tudi po konceptu odzivnega slovarja; gre za odprto dostopno zbirko relevantnih, a še ne povsem neprečiščenih podatkov. Pri izdelavi prečiščene baze sodeluje jezikovna skupnost, s čimer izdelava slovarja ni nikdar zaključena, saj se soustvarja skladno s spremenljivo jezikovno realnostjo. Poleg soustvarjanja jezikovni uporabniki potencialne iztočnice tudi vrednotijo s svojimi odzivi (Arhar Holdt et

al., 2018). Uporabniki prednost koncepta prepoznavajo v preglednosti, dostopnosti, hitremu prilagajanju sodobni sliki jezika, soustvarjalnosti, preprosti uporabi in načinu razvrščanja iztočnic (Kojc et al., 2018; Kamenšek Krajnc et al., 2018). Temu bi moral slediti tudi sodobni slovar protipomenk..

3. Metodologija

3.1. Pridobivanje podatkov

Podatkovno množico s protipomenkami smo sestavili iz več virov. Postopek je podrobneje opisan v diplomskem delu (Pegan, 2019), z izjemo zadnjega koraka z brisanjem ponavljajočih zapisov, ki je bil dodan naknadno. Glavnino podatkov o protipomenkah smo pridobili iz baze sloWNet (Fišer, 2015), manjši delček (87) pa na osnovi klicev iz slovarja SSKJ, dostopnega na slovarskem portalu Fran. Baza sloWNet ima obliko XML, poglejmo si en primer zapisa množice sopomenk (*synset*):

```
<SYNSET>
    <ID>eng-30-00001740-a</ID>
    ...
    <SYNONYM xml:lang="en">
        <LITERAL sense="1"
                pwnid="able%3:00:00::">able
        </LITERAL>
    </SYNONYM>
    <SYNONYM xml:lang="sl">
        <LITERAL lnote="auto">sposoben</LITERAL>
        <LITERAL lnote="auto">zmožen</LITERAL>
    </SYNONYM>
    ...
    <ILR type="near_antonym">
        eng-30-00002098-a</ILR>
    ...
</SYNSET>
```

Za vsak *synset* smo poiskali protipomenski *synset* prek elementa 'near_antonym'. Uporabili smo vse kombinacije, kjer je ena beseda v izvornem *synsetu* in druga v protipomenskem *synsetu*. Na tak način smo pridobili 4.514 parov protipomenk.

Iz SSKJ smo poiskali vsa gesla, ki imajo navedene tudi protipomenke. Poenostavljen primer zapisa vidimo spodaj:

```
<div>
    <span title="Iztočnica">abstrákten</span>
    ...
    <span title="Protipomenka">ant. </span>
    <span title="Protipomenka">
        <a>konkreten</a>:</span>
    ...
</div>
```

Skupno smo iz SSKJ izluščili 87 parov protipomenk.

Zaradi maloštevilčnosti smo podatke o protipomenkah razširili tako, da smo dodajali pare besed s pripono ne-, proti-, brez-. Primeri tako pridobljenih parov so *dostopen – nedostopen*, *ustaven – protiustaven* ter *alkoholen – brezalkoholen.* Tako pridobljene podatke smo deloma ročno prečistili nesmiselnih kombinacij, kot je *no – brezno* ter odstranili besede, za katere nismo imeli vektorskih vložitev v okviru diplomske naloge. Tako smo dobili 1340 parov protipomenk. Dodatno smo upoštevali tudi pare protipomenk, kjer eno izmed obeh besed zamenjamo z njeno sopomenko, s čimer se je množica povečala na 4113 parov protipomenk. Po brisanju ponavljajočih se zapisov, kjer sta besedi le zamenjani, smo pridobili množico 2852 parov protipomenk.

3.2. Označevanje podatkov

V raziskavo je bilo vključenih 2852 parov protipomenk. Vsak izmed šestih pregledovalcev je pregledal vse primere v individualni Google Preglednici, pri čemer je vsakemu paru pripisal eno izmed možnosti d, g in n. Oznaka d označuje, da gre za protipomenki, oznaka n pove, da dani besedi nista protipomenki, oznaka g pa pomeni, da je par problematičen in ga je treba podrobneje proučiti. Označevalci pred začetkom nismo prejeli natančnejših navodil, kaj se smatra kot protipomensko in kaj ne. Namen prvega koraka je bil namreč na osnovi gradiva ugotoviti problematična področja, ki bi jih lahko podrobneje analizirali v nadaljevanju.

Med pregledovanjem smo beležili primere in sproti oblikovali 19 problemskih kategorij. V nadaljevanju smo vsakemu izmed problematičnih parov pripisali po en glavni in morebitni dodatni problem. Podatke smo si razdelili na tri dele, pri čemer je vsak pregledovalec pregledal dva dela podatkov. Med pregledovanjem smo dodali še dve novi kategoriji, in sicer *(Ne)dovršne glagolske tvorjenke* in *Dejanje in stanje*, saj sta se kot problematični izkazali šele po natančnejši analizi vseh primerov.

4. Rezultati in analiza

Po prvem krogu pregledovanj smo 1124 (39,4 %) parov enotno potrdili kot protipomenske in le 22 (0,8 %) primerov kot neprotipomenske. Pri preostalih (1706; 59,8 %) se je vsaj eden izmed pregledovalcev odločil drugače kot ostali, zato smo takšne primere označili za nadaljnjo analizo. V drugem krogu pregledovanja pa se je izkazalo, da so bili nekateri primeri problematični zgolj v zelo specifičnem pogledu oz. da je bil primer lažno označen kot problematičen. Odločitev je bilo treba spremeniti tudi pri nekaterih že potrjenih parih, saj so se po podrobnejšem pregledu izkazali kot problematični. Kategorija potrjenih protipomenk se je tako povečala na 1207 (42,3 %) primerov, medtem ko je bilo potrjenih neprotipomenskih parov 48 (1,7 %). V nadaljnjo analizo smo poslali 1597 (56 %) primerov, kot prikazuje Tabela 1.

Oznaka	Delež
Sta protipomenki	42,3 %
Nista protipomenki	1,7 %
Nadaljnji pregled	56,0 %

Tabela 1: Rezultati po drugem krogu označevanja.

Nadaljnja raziskava se bo osredotočila zgolj na primere (1597; 56 %), ki so se po drugem krogu pregledovanj izkazali za problematične. Razdelili smo jih v 21 kategorij, prikazanih v Tabeli 2, kjer smo za lažjo predstavo vsako izmed kategorij ponazorili s primerom para besed, o katerih smo presojali. Vidimo lahko tudi, kolikokrat se je vsaka izmed kategorij pojavila kot glavni in kot dodatni problem. Glavne probleme smo določili 1597 primerom, medtem ko

smo dodatni problem identificirali pri 668 (41,83 %) primerih, ki predstavljajo 23,46 % celotnega gradiva.

Iz Tabele 2 je razvidno, da se kot glavni problem najpogosteje pojavlja *Redkost in kontekstualna vezanost* pomenov (31,87 %). Pogosto se pojavljajo tudi kategorije *Zanikanost s predpono -ne in -brez* (10,58 %), *Nedoslednost na ravni prevzeto – podomačeno* (10,33 %) in *Zaznamovanost in/ali redkost besede* (9,83 %). Najredkeje so se kot problematične pojavljale kategorije *Zatipki* (0,31 %), *Drugo* (0,38 %) in *Pomensko šibki glagoli* (0,44 %).

Kategorija	Primer	Št. pojavitev (glavni problem)	Odstotek	Št. pojavitev (dodatni problem)	Odstotek
Zatipki	*čistost – nečistot*	5	0,31	/	/
Napačne leme	*alkoholne – brezalkoholne*	40	2,50	3	0,45
Različna besedna vrsta	*dopoldne – popoldanski*	16	1,00	/	/
(Ne)dovršnost	*narasti – zniževati*	87	5,45	2	0,30
(Ne)določnost	*bližnji – daljen*	11	0,69	/	/
Neobstoječe besedotvorne različice	*pritrjevanje – zanikanost*	54	3,38	7	1,05
Zanikanost s predpono ne, brez-	*občutljivost – nedražljivost*	169	10,58	201	30,09
Nedoslednost na ravni prevzeto - podomačeno	*aktiv – trpnik*	165	10,33	36	5,39
(Ne)dovršne izglagolske tvorjenke	*zmanjšanje – povečanje*	32	2,00	4	0,60
Dejanje in stanje	*brezposelnost – zaposlitev*	18	1,13	2	0,30
Povratnost	*ubogati – upirati (se)*	53	3,32	17	2,54
Pomensko šibki glagoli	*manjkati – biti (prisoten)*	7	0,44	2	0,30
Pomensko polne besede	*pridobiti – odreči (soglasje)*	15	0,94	2	0,30
Spol kot “protipomenka”	*kralj – kraljica; dolžnica – upnik*	60	3,76	3	0,45
Zaznamovanost in/ali redkost besede	*ata – mati; nenavadno – često*	157	9,83	79	11,83
Enakopisnice in večpomenke	*bistrost – motnost*	76	4,76	20	2,99
Redkost in kontekstualna vezanost pomenov	*bogat – neploden*	509	31,87	246	36,83
Lastnosti, ki si niso protipomenske, a se pogosto tako uporabljajo	*krivulja – premica*	38	2,38	11	1,65
Posredne sopomenke	*glasen – nem*	40	2,50	5	0,75
Stopenjski primeri	*prihodnji – sedanji*	39	2,44	17	2,54
Drugo	*ofenziven – nespotakljiv*	6	0,38	11	1,65

Tabela 2: 21 kategorij in njihove pojavitve kot glavni in dodatni problem.

4.1. Zatipki

V kategorijo *Zatipki* spadajo pari, pri katerih je vsaj ena izmed besed nedvoumno zatipkana, torej ne more biti isti ali drug leksem v katerikoli obliki. Iz Tabele 2 je razvidno, da se je ta kategorija pojavila zgolj petkrat (0,31 %) kot glavni problem in nikoli kot dodatni. Je ena izmed najbolj problematičnih kategorij, saj besed, ki so narobe črkovane, ne moremo vključiti v slovar.

Primeri: *čistost – nečistot, izginti – pojaviti, izvažati – uvžati*..

4.2. Napačne leme

Pod *Napačne leme* sodijo primeri, ki so sicer lahko oblikoslovno ujemajoči, vendar v neslovarski obliki. Iz Tabele 2 je razvidno, da se je ta kategorija pojavila v 40 primerih (2,5 %) kot glavni in trikrat (0,45 %) kot dodatni problem. Takšne primere moramo odstraniti s seznama parov za vključitev v slovar oz. jih spremeniti v pravo slovarsko obliko.

Primeri: *alkoholne – brezalkoholne, dolžna – nedolžna, finančne – nefinančne.*

4.3. Različna besedna vrsta

Pri kategoriji *Različna besedna vrsta* gre za besedne pare, kjer sestavini pripadata različnima besednima vrstama (npr. samostalnik in pridevnik, pridevnik in prislov). Kot glavni problem se je ta kategorija pojavila pri 16 parih (1,00 %), kot sekundarni pa sploh ne. Pri večini primerov besedi nista protipomenki, dilema se pojavi le pri parih tipa samostalnik – pridevnik, saj gre tukaj največkrat za posamostaljene pridevnike (tipa *delavnik – fraj*). V takšnih primerih sta besedi lahko rabljeni protipomensko, seveda v ustreznem kontekstu. Pare iz te kategorije se odstrani s seznama za vključitev v slovar. Izjemo predstavljajo pari tipa samostalnik – pridevnik, ki se jih ročno pregleda in vključi s potrebnimi oznakami.

Primeri: *dopoldne – popoldanski, znotraj – ven, delavnik – fraj.*

4.4. (Ne)dovršnost

Pri *(Ne)dovršnosti* govorimo o glagolskih parih z različnim glagolskih vidom. Tako je eden izmed glagolov v nedovršni, drugi pa v dovršni obliki. Takšni pari so bili v 87 (5,45 %) primerih prepoznani kot primarni in dvakrat (0,30 %) kot sekundarni problem. Jasno je, da je za protipomenko nekemu glagolu najboljša izbira glagol, ki ima enak glagolski vid, a dilema ostaja pri glagolih, ki so pomensko ustrezni in imajo drugačen glagolski vid. Takšne pare bi bilo (vsaj na prvi pogled) smiselno odstraniti.

Primeri: *napasti – braniti, narasti – zniževati, natovoriti – iztovarjati.*

4.5. (Ne)določnost

V to kategorijo sodijo pridevniški pari, pri katerih je eden izmed pridevnikov v določni, drugi pa se pojavlja v nedoločni obliki. Ta kategorija se je v 11 (0,69 %) primerih pojavila kot glavni problem, medtem ko se kot dodatni problem ni pojavila. Ker je problem v veliki meri povezan z značilnostmi lematizacije za slovenščino, ki pridevnike lematizira v nedoločno obliko, razen kadar to ni mogoče (pari so pomensko načeloma protipomenski), bi bilo tovrstno gradivo smiselno ohraniti v slovarju.

Primeri: *bližnji – daljen, mesten – podeželski, oddaljen – bližnji.*

4.6. Neobstoječe besedotvorne različice

Gre za primere, ki so pomensko sicer ustrezni, a se težava pojavi, ker je ena (ali obe) beseda(i) neobstoječa. Kot primarni problem se je ta kategorija pojavila v 54 (3,28 %) primerih, kot sekundarni pa v sedmih (1,05 %). Ta kategorija po naši presoji ne sodi v slovar, saj gre za besede, ki niso realno v rabi. Že pri luščenju protipomenskih kandidatov bi lahko dodali korak preverbe posamezne besede v referenčnem korpusu in dodali opozorilo pri tistih primerih, ki se ne pojavljajo.

Primeri: *pritrjevanje – zanikanost, eleganca – neelegantnost, nelaskav – podrepniški.*

4.7. Zanikanost s predpono ne-, brez-

V kategoriji *Zanikanost s predpono -ne, -brez* govorimo o primerih, pri katerih je vsaj ena izmed protipomenk tvorjena kot negacija nekega izraza. Gre za pare, kjer sta obe besedi negaciji dveh protipomenk ali za primere, kjer se kot protipomenski par pojavita beseda in negacija njene sopomenke. Kot je razvidno iz Tabele 2, je bila ta kategorija v 169 (10,58 %) primerih prepoznana kot glavni in v 201 (30,09 %) kot dodatni problem. Raba takšnih parov v besedilu bi bila morda slogovno problematična, zagotovo pa so protipomenski v določenih kontekstih. Pare bi zato vključili v slovar in odločitev prepustili uporabniku, ki najbolje pozna kontekst, v katerem se beseda nahaja.

Primeri: *nespremenljiv – nestalen, neugoden – škodljiv, koristen – neugoden.*

4.8. Nedoslednost na ravni prevzeto – podomačeno

Tukaj obravnavamo primere, ki so sicer protipomenski, a je ena izmed besed prevzeta in s tem pogosto (drugače) zaznamovana. Zanimivo je tudi iskati mejo med »prevzetim« izrazom (*ujemanje – inkongruenca*) in takim, ki je v jeziku že uveljavljen (*inteligenten – neumen*). Razlike se lahko pojavljajo tudi na ravni zapisa prevzete besede in ne le v njenem pomenu (npr. *software* in *softver*). Iz Tabele 2 je razvidno, da so označevalci vsaj eno besedo prepoznali kot prevzeto v 165 (10,33 %) primerih, kjer je bil to glavni problem in v 26 (5,39 %) primerih, kjer je bil to dodatni problem. Ker gre tu le za prevzete besede, ki se v jeziku (še) niso uveljavile, bi jih bilo dobro vključiti v odzivni slovar, saj jih bo uporabnik lahko s pridom uporabljal v primernih kontekstih.

Primeri: *aktiv – trpnik, politeizem – enoboštvo, skupen – individualen.*

4.9. (Ne)dovršne glagolske tvorjenke

V kategorijo *(Ne)dovršne glagolske tvorjenke* sodijo tvorjenke, pri katerih besedotvorna podstava izkazuje razlike v dovršnosti. Ena beseda je torej tvorjena iz dovršnega, druga pa iz nedovršnega glagola. Analiza je pokazala, da je primerov, kjer je bila ta kategorija prepoznana kot primarni problem, 32 (2 %), in da so takšni, kjer je bila prepoznana kot sekundarni, štirje (0,60 %). Ti primeri so podobni 4.4, zato bi jih bilo smiselno obravnavati na enak način, torej jih ne bi vključili v slovar.

Primeri: *zmanjševanje – povečanje, izkrcanje – vkrcavanje, manjšanje – povečevanje.*

4.10. Dejanje in stanje

V to kategorijo sodijo samostalniški pari, ki so sicer protipomenski, a ena beseda predstavlja neko dejanje, dogodek, drugi pa neko stanje, lastnost. Problematika je podobna kot pri *(Ne)dovršnih glagolskih tvorjenkah*, le da gre tu za samostalnike, ki niso glagolsko tvorjeni. Kot je razvidno iz Tabele 2, se je kategorija *Dejanje in stanje* kot glavni problem pojavila v 18 (1,13 %) primerih in kot dodatni v 2 (0,30 %) primerih. Ker gre pri takšnih parih za manjšo nianso v pomenu, ki so v določenih kontekstih lahko protipomenski, jih je najbolje uvrstiti v slovar in uporabniku omogočiti, da sam presoja o njihovi uporabnosti.

Primeri: *zaposlitev – brezposelnost, degeneracija – razvoj, nedolžnost – zagrešitev.*

4.11. Povratnost

V kategorijo *Povratnost* smo uvrstili glagolske pare, ki so sicer protipomenski, a vsaj enemu izmed njiju (ali obema) manjka povratni zaimek. Brez povratnega zaimka takšni glagoli nimajo smisla ali imajo drugačen pomen (ki ni protipomenski s predlagano protipomenko). Iz Tabele 2 je razvidno, da je povratni zaimek kot glavni problem manjkal v 53 (3,32 %) parih, in pri 17 (2,54 %) kot dodatni problem. Ker je pri takšnih glagolih povratni zaimek ključen za smiselnost protipomenskega para, ga je nujno dodati. Takšne primere bi zato odstranili s seznama za vključitev v slovar.

Primeri: *strinjati (se) – prepirati (se), ubogati – upirati (se), udeležiti (se) – zamuditi.*

4.12. Pomensko šibki glagoli

Pri pomensko šibkih glagolih govorimo o glagolskih parih, v katerih (vsaj) en člen ob sebi zahteva dopolnilo, če ga želimo smatrati kot protipomenko drugemu. Kategorija se je kot glavni problem pojavila sedemkrat (0,44 %) in dvakrat (0,30 %) kot dodatni. Če naj bodo tovrstni primeri uvrščeni v slovar, mora biti ob pomensko šibkem glagolu dodana ustrezna beseda ali zveza.

Primer: *manjkati – biti (prisoten), biti (statičen/pri miru) – premikati (se), biti (statičen/pri miru) – gibati (se).*

4.13. Pomensko polne besede brez konteksta

Pod *Pomensko polne besede* spadajo pari, kjer je en člen lahko uporabljen kot protipomenka drugemu le takrat, ko je uporabljen v določenem kontekstu skupaj z neko drugo besedo. V ostalih kontekstih besedi nista v protipomenskem razmerju. Kot glavni problem se je omenjena kategorija pojavila pri 15 (0,94 %) parih in kot dodatni pri 2 (0,30 %) parih. Zdi se, da bi tovrstne probleme v slovar lahko vključili, manko konteksta pa rešili na ravni kolokacij, ki jih Slovar sopomenk sodobne slovenščine trenutno vključuje za pomensko primerjavo dveh sopomenk.

Primeri: *pridobiti – odreči (soglasje), odpovedati – obdržati (naročnino), napolniti – sprožiti (pištolo).*

4.14. Spol kot »protipomenka«

V kategoriji *Spol kot »protipomenka«* sta se pojavljali dve problematiki. Najprej smo obravnavali pare, kjer sta kot protipomenki navedena izraza, ki ju uporabljamo za označevanje spolov. Vprašanje je, ali je ob upoštevanju želene družbene občutljivosti slovarja spol sploh ustrezno definirati kot »protipomenski« ter ga s tem obravnavati kot nekaj nasprotnega in binarnega (npr. *moški – ženska*). Druga problematika je razvidna tudi iz primerov, pri katerih sta bila samostalnika (tipično) protipomenska, a v različnih slovničnih spolih (*dolžnica ¬ upnik*). Kot glavni problem se je spol pojavil pri 60 (3,76 %) parih in kot dodatni pri 3 (0,45 %) parih. Če bi kategorijo kljub problematičnosti uvrstili v slovar, bi bilo smiselno natančneje opazovati odzive uporabnikov in ugotoviti, kako ocenjujejo uporabnost in primernost tovrstnega gradiva. Pari tipa *dolžnica – upnik* niso ustrezni za v slovar oz. bi treba gradivo umestiti pod ustrezno iztočnico (*dolžnica – upnica*; *dolžnik – upnik*).

Primeri: *moški – ženska, kralj – kraljica, dolžnica – upnik.*

4.15. Zaznamovanost in/ali redkost besede

V kategoriji *Zaznamovanost in/ali redkost besede* najdemo pare, kjer načeloma gre za protipomenska izraza, a je en izmed njiju zaznamovan. V nekaterih primerih gre za čustveno zaznamovanost (*fant – punči*), v drugih za zastarelo rabo (*izjemoma – često*), pogovorne izraze (*delavnik – fraj*) ali zgolj za izraze, ki se v rabi le redko pojavljajo (*debelost – mršavost*). Kot je razvidno iz Tabele 2, se je ta kategorija kot glavni problem pojavljala precej pogosto, in sicer pri 157 (9,83 %) parih, prav tako pa tudi kot dodatni problem (pri 79 parih, tj. 11,83 %). Ker gre za primere, ki semantično ustrezajo pojmu protipomenskosti, bi jih bilo najbolje vključiti v slovar, da uporabnik sam preceni, če oz. kdaj so v njegovem kontekstu uporabni. Zagotovo bi jim pa bilo dobro dodati slovarsko oznako, ki bi označevala zaznamovanost, ki jo takšni izrazi imajo.

Primeri: *brat – sestrica, izredno – vobče, dolgovezen – koncizen.*

4.16. Enakopisnice in večpomenke

V kategorijo *Enakopisnice in večpomenke* so vključeni pari, kjer je eden izmed izrazov večpomenski. Pri teh parih gre velikokrat tudi za prenesen pomen enega izmed členov (*hladen – navdušen*). Problematične so tudi prave enakopisnice, torej tiste, ki bi v slovarju imele ločene iztočnice in ne le več pomenov (*pust – masten*). Takšni pari so se kot glavni problem pojavili 76-krat (4,76 %) in 20-krat (2,99 %) kot dodatni problem. Takšni primeri seveda sodijo v slovar, treba pa bi bilo opredeliti, s katerim pomenom besede je določena beseda v protipomenskem razmerju.

Primeri: *bistrost – motnost, zajedalec – gostitelj, moder – naiven.*

4.17. Redkost in kontekstualna vezanost primerov

V to kategorijo sodijo primeri, ki so protipomenski le v določenih kontekstih. Običajno je tu eden izmed izrazov bolj uveljavljen in uporabljen v več kontekstih, zato je protipomenka drugemu le v določenih primerih. Prav tako so se tukaj znašli primeri, pri katerih bi bili sestavini para v nad/podpomenskem razmerju, če bi eno od njiju negirali (kot pri *zdrav – umobolen*, kjer bi bili pravi protipomenki *zdrav – bolan*, medtem ko je *umobolen* le ena oblika nezdravja). V kategorijo *Redkost in kontekstualna vezanost primerov* smo vključili tudi primere, kjer je bil eden izmed izrazov zelo specifičen, običajno terminološki (primer: *izdelava – delaboracija*). Odločili smo se, da terminoloških izrazov ne bomo uvrščali v posebno kategorijo, saj je težko

določiti mejo med strokovnimi, specifičnimi in »čistimi« terminološkimi izrazi. Ker gre za najširšo kategorijo, se je kot primarni problem pojavila v kar 509 (31,87 %) primerih in kot sekundarni v 246 (36,83 %) primerih. Te primere se vključi v odzivni slovar, saj lahko uporabnik v široki paleti možnosti izbere zase najustreznejšo.

Primeri: *bogat – neploden, cena – prednost, domač – nepoznan.*

4.18. Lastnosti, ki niso protipomenske, a se pogosto tako uporabljajo

V tej kategoriji so zbrani primeri, ki sicer opisujejo izključujoče lastnosti, a v strogem pomenu ne gre za protipomenki, čeprav se pogosto tako uporabljata. To so predvsem pari, ki jih v pogovornem kontekstu uporabljamo kot protipomenki, ali takšni, za katere zmotno mislimo, da to sta. Kot je razvidno iz Tabele 2, je bila ta problematika prepoznana v 38 (2,38 %) primerih kot glavni in v 11 (1,65 %) primerih kot dodatni problem. Čeprav takšni pari niso strogo gledano protipomenski, bi jih bilo najverjetneje smiselno vključiti v slovar in izbiro prepustiti uporabniku.

Primeri: *anabolizem – katabolizem, krivulja – premica, nepomemben – znamenit.*

4.19. Posredne sopomenke

Pod *Posredne sopomenke* sodijo pari tipa *glasen – nem*, ki so na prvi pogled protipomenski le v redkih primerih, če pa bi eno sestavino zamenjali z njeno sopomenko, bi dobili precej bolj očiten protipomenski par (npr. *glasen – tih*). Takšni pari so se kot primarni problem pojavili 40-krat (2,50 %), kot sekundarni pa 5-krat (0,75 %). Čeprav niso prototipsko protipomenski, bi bilo tudi takšne pare morda dobro vključiti v slovar, saj uporabniku lahko koristijo v določenih situacijah, obenem pa spremljati, ali bodo uporabniki v odzivnem slovarju tovrstne primere ocenjevali s pozitivnimi ali negativnimi glasovi.

Primeri: *profit – minus, glasen – nem, kvaren – koristen.*

4.20. Stopenjski primeri

V to kategorijo smo zbrali pare, ki jih sicer lahko razumemo kot protipomenske v določenem kontekstu, a se pojavlja zelo očitna stopnjevanost. Besedi torej sta lahko protipomenki (*prihodnji – sedanji*), a običajno obstaja še neko bolj izrazito nasprotje (*prihodnji – pretekli*).Sem smo vključili tudi stopnjevane pridevnike, ki pa niso vedno nujno na popolnoma nasprotni stopnji. Tako imamo lahko v paru npr. primernik in presežnik in ne le dva primernika (primer: *manjši – največji* in ne le *manjši – večji*). Stopenjski primeri so se kot glavni problem pojavili v 39 (2,44 %) primerih in v 17 (2,54 %) primerih kot dodatni problem. Ker so kontekstualno pogojeni, jih je dobro vključiti v odzivni slovar in tako uporabniku omogočiti širšo izbiro potencialnih protipomenk.

Primeri: *negativen – nevtralen, dvojen – enojen, maksimalen – majhen*.

4.21. Drugo

Pod *Drugo* smo vključili primere, ki niso sodili v nobeno izmed ostalih kategorij. Kot je razvidno iz Tabele 2, smo 6 (0,38 %) parov vključili pod Drugo kot glavni problem in 11 (1,65 %) parov kot dodatni problem. Takšne pare, ki so se pojavili zelo poredko (0,38 %), bi bilo smiselno vključiti v slovar in presojo uporabnosti prepustiti uporabniški skupnosti.

Primeri: *državljan – tujec, ofenziven – nespotakljiv, zamuditi – zadeti.*

5. Zaključek

Iz analize je razvidno, da imajo problemske kategorije različno težo, nekatere težave bi bilo treba nasloviti, preden se gradivo lahko vključi v slovar, medtem ko lahko pri drugih odločitev o relevantnosti prepustimo uporabniški skupnosti. V analizi smo ugotovili, da so kategorije *Zatipki*, *Napačne leme*, *Različna besedna vrsta*, *(Ne)dovršnost*, *Neobstoječe besedotvorne različice*, *(Ne)dovršne glagolske tvorjenke* in *Povratnost* najbolj problematične, vendar jih je obenem predvidoma mogoče vsaj delno reševati tudi avtomatsko, kar bomo upoštevali pri razvoju nadaljnje metodologije strojnega pridobivanja protipomenk. Ostale kategorije pa so bolj vezane na kontekst, zato jih lahko vključimo v slovar in odločitev prepustimo skupnosti.

Čeprav je bilo nedvoumno potrjenih protipomenk na prvi pogled malo (manj kot polovica), pa nadaljnja analiza kaže, da lahko v odzivni slovar vključimo veliko večino (88 %) podatkov. Prav tu se kaže prednost odzivnega slovarja, ki uporabniku ponuja možnost, da izbira med širokim naborom potencialnih protipomenk in jih ocenjuje kot bolj ali manj ustrezne. V slovar je torej najbolje vključiti čim več potencialnega gradiva in jezikovni skupnosti prepustiti odločitev, kaj je zanjo uporabno in kaj ne.

Z digitalizacijo družbe so se spremenile (in povečale) potrebe jezikovnih uporabnikov, ki želijo vedno večji nabor podatkov, med katerimi lahko izbirajo. Odzivni slovar jim ne omogoči zgolj tega, ampak tudi dodajanje novega gradiva in odzivanje na že obstoječe. Skupaj z družbo se tako spreminjajo slovarji, z njimi pa tudi mi in naša vloga pri njihovem ustvarjanju.

6. Zahvala

Projekt *Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL* v letih 2021–22 financira Ministrstvo za kulturo Republike Slovenije.

Avtorji in avtorice bi se radi zahvalili tudi Špeli Arhar Holdt za vključitev v projekt in pomoč pri načrtovanju raziskave in prispevka.

7. Literatura

Luluh Aldhubayi in Maha Alyahya. 2014. Automated Arabic Antonym Extraction Using a Corpus Analysis Tool. *Journal of Theoretical and Applied Information Technology,* 70(3):422–433.

Darja Fišer. 2015. Semantic lexicon of Slovene sloWNet 3.1. *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1026.

Polona Gantar, Iztok Kosem in Simon Krek. 2016. Discovering automated lexicography: the case of the slovene lexical database. *International Journal of Lexicography*, 29(2):200–225.

Špela Arhar Holdt, Jaka Čibelj, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski in Marko Robnik Šikonja. 2018. Thesaurus of Modern Slovene: By the Community for the Community. V: *Thesaurus of Modern Slovene: By the Community for the Community.*

Proceedings of the XVIII EURALEX International Congress, str. 401–410.

Marjeta Humar. 2005. *Protipomenskost v slovenski jezikoslovni literaturi*. V: M. Jesenček, ur., *Knjižno in narečno besedoslovje slovenskega jezika,* str. 234–238, Slavistično društvo Maribor, Maribor.

Marjeta Humar. 2016. *Protipomenskost v slovenskem knjižnem jeziku: na primeru terminoloških slovarjev*. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Ljubljana.

Elin Kamenšek Kranjc, Špela Medved in Kaja Podgoršek. 2018. Primerjava spletnega slovarja Slovar sopomenk sodobne slovenščine in knjižnega Sinonimnega slovarja slovenskega jezika. *Liter jezika,* 9(12):66–70.

Agnes Kojc, Tamara Rigler, Kaja Sluga, Anika Plešivčnik in Špela Kovačič. 2018. Slovar sopomenk sodobne slovenščine in Sinonimni slovar slovenskega jezika. *Liter jezika,* 9(12):62–65.

Simon Krek, Cyprian Laskowski, Marko Robnik Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemec in Kaja Dobrovoljc. 2018. Thesaurus of Modern Slovene 1.0. *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1166.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem in Kaja Dobrovoljc. 2020. Gigafida 2.0: the reference corpus of written standard Slovene. V: N. Calzolari, ur., *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*, str. 3340–3345. ELRA - European Language Resources Association, Paris. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Nikola Ljubešić in Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1204.

Anna Lobanova, Tom van der Kleij in Jennifer Spenader. 2010. Defining Antonymy: A Corpus-based Study of Opposites by Lexico-syntactic Patterns. *International Journal of Lexicography,* 23(1):19–53.

Ada Vidovič Muha. 2005. *Medleksemski pomenski razmerji – sopomenskost in protipomenskost*. V: M. Jesenšek, ur., *Knjižno in narečno besedoslovje slovenskega jezika*, str. 206–221. Slavistično društvo Maribor, Maribor.

Ada Vidovič Muha. 2021. *Slovensko leksikalno pomenoslovje. Prva e-izdaja.* Znanstvena založba FFUL, Ljubljana.

Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja. 2014. Cankarjeva založba, Ljubljana.

Sopomenke 1.0. O slovarju. Center za jezikovne vire in tehnologije. https://viri.cjvt.si/sopomenke/slv/about.

Irena Breznik Stramljič. 2010. *Tvorjenke slovenskega jezika med slovarjem in besedilom.* Mednarodna založba Oddelka za slovanske jezike in književnosti FFUM, Maribor.

Jasmina Pegan. 2019. *Detekcija antonimov z vektorskimi vložitvami besed.* Diplomsko delo. Fakulteta za računalništvo in informatiko Univerze v Ljubljani.

Jože Toporišič. 1976. *Slovenska slovnica*. Založba »Obzorja«, Maribor.

Jože Toporišič. 2000. *Slovenska slovnica. Četrta, prenovljena izdaja.* Založba »Obzorja«, Maribor.

Wenbo Wang, Christopher Thomas in Amit Sheth. 2010. Pattern-Based Synonym and Antonym Extraction. *ACM SE '10: Proceedings of the 48th Annual Southeast Regional Conference*: 1–4. https://dl.acm.org/doi/abs/10.1145/1900008.1900094.