Luščenje terminoloških kandidatov za slovar odnosov z javnostmi

Nataša Logar Berginc*, Špela Vintar, Špela Arhar Holdt*****

* Univerza v Ljubljani, Fakulteta za družbene vede
Kardeljeva ploščad 5, SI-1000 Ljubljana
natasa.logar@fdv.uni-lj.si
** Univerza v Ljubljani, Filozofska fakulteta
Aškerčeva 2, SI-1000 Ljubljana
spela.vintar@ff.uni-lj.si
*** Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, SI-4220 Škofja Loka
spela.arhar@trojina.si

Povzetek

V prispevku prikazujemo analizo luščenja eno- in večbesednih terminoloških kandidatov, ki smo ga izvedli za potrebe priprave slovarja odnosov z javnostmi na podlagi korpusa KoRP z luščilnikom LUIZ. Podrobneje se posvečamo dvojemu: (a) izluščenim enobesednim samostalniškim terminološkim kandidatom, katerih seznam primerjamo s pogostostnim seznamom samostalnikov v KoRP in vrednotimo glede na terminološkost, kot sta jo prepoznala dva področna strokovnjaka, ter (b) izluščenim večbesednim nizom z glagolskim jedrom. Nadgrajeno metodo luščenja in izboljšan prikaz rezultatov smo dopolnili še z analizo priklica. Potrdili oz. ugotovili smo, da je v primerjavi s pogostostnim seznamom terminološki potencial enot v zgornjem delu liste izluščenih samostalnikov večji in da imajo izluščeni glagolski besedni nizi predvsem kolokacijsko vrednost, ne pa tudi terminološke. Analiza priklica je pokazala predvsem nizko stopnjo strinjanja med obema področnima strokovnjakoma, sicer pa je bil priklic razmeroma visok.

Term candidate extraction for the dictionary of public relations

The article describes an analysis of automatic term recognition results performed for single- and multi-word terms with the LUIZ term extraction system. The target application of the results is a dictionary of Public Relations and the main resource the KoRP Public Relations Corpus. Our analysis is focused on two segments: (a) single-word noun term candidates, which we compare with the frequency list of nouns from KoRP and evaluate termhood on the basis of the judgements of two domain experts, and (b) multi-word term candidates with verb as headword. In order to better assess the performance of the system and the soundness of our approach we also performed an analysis of recall. Our results show that the terminological relevance of extracted nouns is indeed higher than that of merely frequent nouns, and that verbal phrases only rarely count as proper terms. The analysis of recall shows low inter-annotator agreement, but nevertheless very satisfactory recall levels.

1. Uvod

Korpusno jezikoslovje je z možnostjo meritev zelo različnih lastnosti jezika prineslo opazen premik zlasti v leksikografiji – pa tudi terminografiji (npr. Atkins, Rundell, 2008; Bergenholtz, Tarp, 1995; Biber, Conrad, Reppen, 1998; Čermák, 2011; Halliday et al., 2004; Hanks, 2008; Leech, 1992; McEnery, Wilson, 1996; Pearson, 1998; Schryver, 2003; Sinclair, 2004; Teubert, Krishnamurty, 2007; pri nas npr.: Krek, 2003; Gorjanc, 2005; Gorjanc, Krek, 2005; Gantar, 2007; Gantar, 2009; Logar, 2007; Vintar, 2008). Na osnovi korpusa, ki ima razviden namen, merila gradnje, odločitve snovalcev ter besedilnozvrstno, časovno in drugo zgradbo, lahko področni strokovnjaki in jezikoslovci, ki sodelujejo pri pripravi terminoloških slovarjev ali terminoloških podatkovnih zbirk, svojo interpretacijo, ki bo posledično podana kot jezikovni opis in verjetno tudi predpis, oprejo na podatke, pridobljene iz resnične jezikovne rabe; z umikanjem nuje po introspekciji kot edini možnosti presoje pa se povečata kredibilnost in veljavnost prepoznavanja vseh pojavov – tudi jezikovnih. Pri takem pristopu je ena ključnih prednosti, ki kvalitativno in kvantitativno presega nekorpusno terminografijo, računalniško podprto luščenje terminoloških kandidatov.

Pristopov k luščenju terminoloških kandidatov je več, pri skoraj vseh pa gre za kombinacijo jezikoslovnega znanja o naravi terminov ter izrabo matematičnih lastnosti porazdelitve besed in besednih nizov v korpusih (Vintar, 2008: 100; Vintar, 2009: 346–347 in tam navedena literatura). Za slovenščino je bilo luščenje terminoloških kandidatov preizkušeno že večkrat, npr. za projekt Voice Tran I in II na področju vojaške terminologije ter na področju računalništva in informatike za Islovar (Vintar, Erjavec, 2008; Vintar, 2009; Vintar, 2010: 47); pa tudi pri pripravi slovarja odnosov z javnostmi. Za slednjega smo že predstavili (Logar, Vintar, 2008), da je bilo uspešno zlasti pri pridobivanju večbesednih terminoloških kandidatov, ki je potekalo po kombinaciji statistične in jezikoslovne metode. Na osnovi takrat prepoznanih pomanjkljivosti smo metodo in prikaz rezultatov izboljšali – v nadaljevanju na delu pridobljenih seznamov prikazujemo, na kakšen način in s kakšno uspešnostjo.

2. Eksperiment

Korpus, iz katerega smo luščili terminološke kandidate odnosov z javnostmi, je korpus KoRP. KoRP vsebuje 1,824.699 pojavnic, je enojezični, sinhroni in trenutno statični korpus strokovnih besedil. Od julija 2007 je prosto dostopen na <http://www.korp.fdv.uni-lj.si/>, iskanje po njem pa trenutno poteka še s Konkordančnikom ASP32. Pred tokratnim luščenjem smo ga ponovno označili z najnovejšo različico označevalnika podjetja Amebis, d. o. o. (Romih, Holozan, 2002; Holozan, 2006; Arhar Holdt, 2011: 22–23, 28–29).

S pomočjo luščilnika LUIZ (Vintar, 2010) smo želeli pridobiti dvoje:

a) enobesedne terminološke kandidate: samostalnike, glagole, pridevnike in prislove;

b) večbesedne terminološke kandidate: samostalniške in glagolske besedne zveze.

Tako eno- kot večbesedne kandidate smo luščili s pomočjo oblikoskladenjskih vzorcev (ti so pri enobesednih enotah sestavljeni le iz enega člena) in terminološke uteži, ki se izračuna na podlagi pogostosti pojavitve v specializiranem korpusu v primerjavi s splošnim korpusom ter frazeološke stabilnosti enote. Skupno smo uporabili 39 oblikoskladenjskih vzorcev, od tega 30 s samostalniškim jedrom, 9 z glagolskim (Tabela 1, prvi stolpec). Za tako obsežno število vzorcev smo se odločili zaradi večje zanesljivosti ocene tovrstnega pristopa k pridobivanju terminov v slovenščini, smo pa že pred začetkom luščenja predvidevali, da bo približno polovica vzorcev dala le malo ali celo nič terminološko zanimivega gradiva.

Ker se med izluščenimi kandidati pogosto znajdejo tudi lastna imena, ki pa za slovarske namene v veliki meri niso zanimiva, smo vse enote, ki so vsebovale besede z veliko začetnico, izločili.

3. Analiza rezultatov

Rezultat luščenja so bili seznami, na katerih je bilo skupno 47.007 večbesednih enot (brez lastnih imen; število po vzorcih in primeri so v Tabeli 1) oz. 16.190 enobesednih enot (brez lastnih imen; Tabela 2).

Vzorec	Število kandidatov	Primer
Samostalniške zveze		
1. P S	17.242	*lokalna skupnost*
2. S S	9.362	*vir informacij*
3. S S S	932	*merilo uspešnosti delovanja*
4. P S S	1.670	*uradni vir informacij*
5. S P S	3.160	*dan odprtih vrat*
6. S D S	4.370	*sporočilo za javnost*
7. S D S S	648	*orodje za doseganje ciljev*
8. S D P S	1.174	*odnos z interno javnostjo*
9. P P S	1.398	*celostna grafična podoba*
10. P D S	621	*vodilni v podjetju*
11. R P S	618	*srednje veliko podjetje*
12. S S S S	53	*dvig kakovosti življenja otrok*
13. S P S S	198	*doseganje poslovnih ciljev organizacije*
14. P S S S	130	*osrednje zanimanje svetovne javnosti*
15. P S P S	381	*refleksivni model komunikacijskega menedžmenta*
16. S S P S	321	*model načrtovanja merljivih ciljev*
17. S P P S	245	*model dvosmernih simetričnih odnosov*
18. P P S S	51	*upravljano komunikacijsko ravnanje organizacije*
19. S S D S	982	*strategija odnosov z javnostmi*
20. P S D S	971	*tržni odnosi z javnostmi*
21. R P S S	51	*vnaprej pripravljen predlog vprašanj*
22. R P D S	63	*tesno povezan s teorijo*
23. P D P S	141	*značilen za blagovno znamko*
24. P D S S	100	*potreben za razrešitev konflikta*
25. R P P S	51	*točno določena ciljna javnost*
26. P S in S	445	*medijski čas in prostor*
27. S S in S	365	*trženje izdelkov in storitev*
28. S in S S	434	*mnenje in stališče javnosti*
29. S in P S	367	*čast in dobro ime*
30. P in P S	463	*tiskani in elektronski mediji*
Glagolske zveze		
31. R G	3.032	*pomembno vplivati*
32. G D S	2.208	*odgovoriti na vprašanje*
33. G R	1.602	*delovati neodvisno*
34. D S G	878	*v nadaljevanju predstavljati*
35. G D R	101	*veljati za učinkovito*
36. G kot S	98	*delovati kot posrednik*
37. G kot P	42	*biti kot nov*
38. D R G	14	*od nekdaj spremljati*
39. G kot R	13	*prepoznati kot pomembno*
SKUPAJ	**47.007**	

Tabela 1: Izluščeni večbesedni terminološki kandidati: število po vzorcih in primeri.

Besedna vrsta	Število kandidatov	Primer
1. S	7.379 (z lastnimi imeni: 10.731)	*javnost, odnos, organizacija*
2. P	4.854	*komunikacijski, blagovni, zaposleni*
3. R	1.379	*veliko, pogosto, vedno*
4. G	2.578	*vplivati, sporočiti, komunicirati*
SKUPAJ	**16.190**	

Tabela 2: Izluščeni enobesedni terminološki kandidati: število po besednih vrstah in primeri.

Sezname smo podrobneje pregledali in tako prišli do ocene relevantnosti metode luščenja, dodatno pa izvedli še analizo priklica.

3.1. Enobesedni terminološki kandidati

Predstavili bomo le analizo vrhnjega dela seznama samostalniških terminoloških kandidatov, ki smo ga ocenjevali z dveh vidikov: (a) v primerjavi s pogostostnim seznamom samostalnikov iz KoRP in (b) glede na oceno terminološkosti, ki sta jo dala dva strokovnjaka s področja odnosov z javnostmi.

a) Že ob primerjalnem ogledu zgolj prvih 20 izluščenih samostalnikov in prvih 20 samostalnikov po pogostosti (Tabela 3) opazimo nekaj razlik: pri luščenju so med prvimi dvajsetimi poimenovanja *management*, *deležnik*, *model* in *novinar* (v Tabeli 3 krepki tisk), ki jih na pogostostnem seznamu med prvimi dvajsetimi ni; nasprotno pa so na pogostostnem seznamu samostalnikov višje, tj. do dvajsetega mesta, uvrščeni *človek*, *skupnost* in *stran* (ležeči tisk).

Terminološko luščenje: prvih 20 samostalnikov	**Pogostost: prvih 20 samostalnikov**
javnost	javnost
odnos	odnos
organizacija	organizacija
komuniciranje	podjetje
podjetje	komuniciranje
medij	medij
znamka	znamka
informacija	leto
leto	informacija
cilj	cilj
okolje	primer
management	skupina
področje	okolje
primer	program
skupina	področje
program	*človek*
deležnik	(zaposleni)*
vloga	vloga
model	*skupnost*
novinar	*stran*

Tabela 3: Vrhnji del seznama izluščenih samostalnikov in samostalnikov po pogostosti v KoRP.
* Pri luščenju med prvimi dvajsetimi, vendar pri pridevnikih.

Pregled razlik, ki so se pokazale do stotega mesta tabele, pokaže, da so pri uporabljenem luščenju višje uvrščeni (torej med prvimi stotimi) *dejavnik*, *izvajanje*, *javnost*, *komunikator*, *manager*, *menedžment*, *tveganje*, *uspešnost*, *vedenje* in *zaupanje*; medtem ko so pri pogostostnem seznamu višje uvrščene besede *kot*, *mesto*, *načrt*, *Publica*, *služba*, *str.*, *svet* in *vlada*. Tudi če pogledamo spremembe mest terminoloških kandidatov znotraj stotega mesta, so premiki zanimivi: *management*, ki je pri terminološkem luščenju na 12. mestu, je po pogostosti na 60. mestu, in tako npr. še *deležnik* – 17. (41.), *oglaševanje* – 29. (44.), *manager* – 31. (130.), *marketing* – 36. (59.), *praktik* – 40. (87.), *načrtovanje* – 53. (81.), *komunikator* – 65. (198.), *uspešnost* – 85. (111.), *tveganje* – 97. (122.). Nazadovali pa so pri luščenju samostalniki *človek* – 29. mesto, po pogostosti pa 17. mesto, *stran* – 33. (20.), *sistem* – 94. (64.), *podatek* – 96. (71.) ipd.

b) Strokovnjaka odnosov z javnostmi sta za potrebe priprave terminološkega slovarja pregledala celotni pogostostni seznam samostalnikov, ki je obsegal skoraj 12.500 enot.[1] Če kot načeloma potrjene s pogostostnega seznama vzamemo le tiste, ki so s strani obeh strokovnjakov dobili kljukico (druge oznake so bile še: *nekaj manjka*, *prečrtano*, *vprašaj*), ugotovimo, da jih je med prvimi dvestotimi takih 117. Če na enak način pogledamo tudi seznam terminološko izluščenih samostalniških kandidatov, je rezultat nekoliko boljši: med prvimi dvestotimi izluščenimi je dvojno kljukico dobilo 125 kandidatov. Če preverimo npr. še mesta od 700 do 800 na obeh seznamih, zopet ugotovimo rahlo prednost terminološko izluščenega seznama, in sicer v razmerju 42 : 38. Pri mestih od 1.000 do 1.100 pa je uporabljeno luščenje glede na odločitve obeh strokovnjakov v še večji prednosti: 41 : 28.

Povzamemo lahko, da med vrhnjim delom seznama samostalnikov, ki se kot terminološki kandidati kažejo po luščenju, in vrhnjim delom pogostostnega seznama samostalnikov ni zelo velikih razlik, a so te z vidika ocene terminološkosti vedno v prid terminološko izluščenim seznamom, kar pomeni, da je pri naboru enobesednih terminov za geslovnik bolj smiselno izhajati iz slednjih.

Mesto na seznamu pri terminološkem luščenju	**Terminološki kandidat**
1.	odnos z javnost
4.	korporativen identiteta
5.	odnos z medij
6.	komunikacijski menedžment
10.	uglednosten kapital
11.	lokalen skupnost
13.	ciljen javnost
15.	krizen management
17.	deležniški skupina
22.	posloven komunikator
25.	komunikacijski aktivnost
27.	odnos z zaposleni
31.	komunikacijski kompetenca
32.	komunikacijski program
33.	komunikacijski management

Tabela 4: Prvih 15 izluščenih večbesednih terminoloških kandidatov (luščenje iz leta 2007), potrjenih s strani dveh strokovnjakov odnosov z javnostmi.

3.2. Večbesedni terminološki kandidati

Uspešnost samodejne ekstrakcije večbesednih terminoloških kandidatov smo s pomočjo dveh strokovnjakov s področja odnosov z javnostmi ocenili že pri prvem luščenju, ki smo ga izvedli leta 2007 in v katerega smo zajeli osem samostalniških oblikoskladenjskih vzorcev (prvih osem v Tabeli 1). Za ponazoritev podajamo vrhnji del takratnega seznama, ki prikazuje terminološke kandidate, za katere sta oba strokovnjaka potrdila, da gre za termine odnosov z javnostmi (Tabela 4). Skupno oceno uspešnosti takratnega luščenja lahko strnemo v ugotovitev, da pri prvih 1.000

[1] Tovrstno ocenjevanje terminološkosti zgolj po seznamu, brez vpogleda v sobesedilo, brez posvetovanja s še drugimi področnimi strokovnjaki, pa tudi brez zelo jasnega predstave o končnem slovarju odpira več vprašanj ter ima omejitve in pomanjkljivosti – pri projektu smo ga izvedli tudi zato, da nanje opozorimo in premislimo alternativne rešitve.

enotah na seznamu "v drugi polovici /.../ sicer narašča delež kolokacij, vendar je na tem seznamu veliko terminov" (Logar, Vintar, 2008: 13).[2]

V nadaljevanju izmed vseh 39 na novo luščenih oblikoskladenjskih vzorcev povzemamo le oceno terminološke zanimivosti vzorcev z glagolom, tj. zvez glagola s prislovom, glagolskih predložnih zvez in zvez glagola s podrednim veznikom *kot* (zadnjih devet vzorcev v Tabeli 1; o enobesednih glagolskih terminih, tudi izluščenih, gl. več v Logar, Vintar, 2008: 8–9, 11).

Analiza je pokazala, da le dva od devetih vzorcev dasta nekaj enot, ki bi jih kot celoto lahko vključili kot samostojno geslo v terminološki slovar odnosov z javnostmi, in da je ob teh dveh vzorcih le še eden, ki ima tovrsten potencial – gre za: *R G* in *G R* ter *G D S*. Natančnejši pregled vseh treh je najprej zajel odstranitev primerov z *biti*, *morati*, *želeti* in *jesti* pri *G D S* (kar je s seznama odstranilo 20 % nerelevantnih primerov) ter pri *R G*/*G R* še *moči*, *hoteti*, *smeti*, *imeti*, *začeti*, *postati*, *dobiti*, *iti*; *lahko*, *tako*, *treba*, *mogoče*, *bolj*, *vedno*, *sam*, *potrebno*, *rad*, *nekaj*, *najbolj*, *zato*, *sicer*, pri *G R* pa še dodatno *veliko*, *glede*, *čim* in *vse*. Na ta način so se pri vzorcu *R G* podatki zmanjšali za polovico, pri *G R* pa za 68 %. Potencialnih glagolskih večbesednih terminov je bilo, kot rečeno, le nekaj, in sicer po naši presoji vsi kot zveza glagola *komunicirati* in prislovov *dvosmerno*, *strateško*, *simetrično*, *javno*, *individualno*, *osebno*, *rutinsko* ter *navzven*. Vse ostalo pri teh treh vzorcih (pa tudi pri drugih, čeprav manj) je tako predvsem upoštevanja vreden prikaz kolokacijskega okolja glagolov in samostalnikov, ki so njihov del – če so ti glagoli in samostalniki terminološki, bodo seveda tudi njihovi besedni nizi ob ustrezni pogostosti kot kolokacije prišli v slovar, sicer pa ne, npr.:

– komunicirati*: komunicirati z/s [ljudmi*, *deležniki*, *javnostmi*, *mediji*, *novinarji*, *okoljem*, *organizacijo*, *potrošniki*, *predstavniki /.../*, *tržiščem*, *vlagatelji]*; *[dobro*, *učinkovito*, *nenehno*, *pravilno*, *premalo*, *prepričljivo*, *uspešno] komunicirati*;

– informacija: *[seznaniti*, *razpolagati*, *početi*, *povezati] z informacijo*; *[iti] za informacijo*; *[temeljiti*, *nanašati se] na informacijo*; *[soditi] med informacije*; *[zaupati] glede informacije*; *[predelati] v informacijo*; *[priti] do informacije*.

Lahko torej povzamemo, da so v našem primeru z luščenjem glagolskih vzorcev nastali seznami, ki z vidika nabora samostojnih iztočnic niso pomembno dopolnilo terminografskega dela, z vidika vsebine slovarja (podatkov znotraj iztočnic) pa to vendarle lahko so – zlasti če predpostavljamo, da želimo v slovar vključiti tudi značilno besedilno okolje terminov. Pri slednjem si je sicer mogoče pomagati tudi z naprednimi korpusnimi orodji, ki tovrstne podatke prikazujejo samodejno, tak je npr. program *Besedne skice/Sketch Engine* (Krek, Kilgarriff, 2006). Uporabnost tega in sorodnih orodij za pripravo terminoloških slovarjev bomo preizkusili v nadaljevanju projekta. Analiza je obenem izpostavila seznam glagolov in prislovov, ki bi jih bilo pri eventualnih prihodnjih luščenjih glagolskih zvez smiselno avtomatsko izločiti iz končnega seznama rezultatov in s tem podatke vnaprej selekcionirati; gre denimo za glagole *biti*, *imeti*, *iti*, *dati*, modalne glagole, deloma fazne glagole ipd. ter del prislovov, predvsem tistih, ki so v jeziku zelo pogosti, pomensko pa zelo splošni (*tako*, *zelo*, *lahko*, *vedno* itd.).

3.3. Priklic

Kot je razvidno iz dosedanje razprave, se je natančnost luščenja pri posameznih vzorcih močno razlikovala, različne evalvacije iz preteklih eksperimentov pa kažejo, da se pri osnovnem naboru oblikoskladenjskih vzorcev natančnost luščenja za prvih 100 kandidatov giblje med 70 in 90 % (Vintar in Erjavec, 2008; Vintar, 2010). A medtem ko je merjenje natančnosti razmeroma enostavno, saj moramo zgolj pregledati vrh seznama predlaganih kandidatov in ugotoviti, koliko jih je pravih terminov, je ugotavljanje priklica bistveno težje.

S priklicem pri jezikovnih tehnologijah merimo sposobnost sistema, da v množici podatkov prepozna zadovoljiv odstotek iskanih primerov. Z drugimi besedami: pri našem eksperimentu nas je zanimalo, koliko terminov, ki se v besedilu pojavijo, ostane neizluščenih. V ta namen smo izbrali znanstveni članek s področja odnosov z javnostmi in prosili dva področna strokovnjaka (ne ista strokovnjaka, kot sta pregledovala sezname), da v članku označita vse terminološko relevantne izraze. Iz obeh člankov smo nato izpisali označene termine in opazovali dvoje:

a) v kolikšni meri se mnenji obeh strokovnjakov o terminološkosti ujemata (Tabela 5);

b) v kolikšni meri se nabor "človeško" izbranih terminov ujema s samodejno izluščenim seznamom (Tabela 6).

Prvi strokovnjak	184
Drugi strokovnjak	261
Presek	109
Unija	415
Ujemanje med strokovnjakoma (IAA)	0,26

Tabela 5: Ujemanje pri oceni terminološkosti med obema strokovnjakoma.

Kot vidimo, sta strokovnjaka v istem besedilu označila zelo različno število terminov, saj je presek med njima le 109. Če njuna izbora združimo, dobimo unijo v velikosti 415 terminov, v nadaljevanju pa merimo priklic na obeh seznamih.

	Priklic na preseku	**Priklic na uniji**
Vsi izluščeni terminološki kandidati (63.179)	0,93	0,85
Prvih 10.000	0,84	0,72
Prvih 5.000	0,75	0,63

Tabela 6: Priklic izluščenih terminoloških kandidatov.

[2] Pri tem smo ločili med (a) večbesednim terminom kot stalno zvezo, poimenovanjem in (b) kolokacijo kot leksikalno in/ali pragmatično povezano ponovljivo sopojavitvijo vsaj dveh leksikalnih enot, ki sta med seboj v neposrednem skladenjskem razmerju (Bartsch, 2004, nav. po Heid, 2006: 980), a v ožjem pomenu, tj. kot prosto zvezo. Razmejitev je seveda groba (prim. Vintar, 2003: 74; Erjavec, Vintar, 2004: 104; Logar, Vintar, 2008: 12–14), jo pa ohranjamo tudi v tem prispevku.

Rezultati priklica so dobri. Od 109 terminov ki sta jih označila oba strokovnjaka, jih sistem le 7 ni izluščil in sicer gre v treh primerih za angleške izraze (*issue management*, *press clipping*, *cluster analiza*), v dveh primerih za petbesedne enote, naši vzorci pa ne presegajo dolžine štirih besed (*pragmatična raven odnosov z javnostmi*, *odzivni razvojni model strateškega načrtovanja*), in v enem primeru za vzorec *P P*, česar pri vzorcih nismo predvidevali (*dvosmerni asimetrični*). Kot je pričakovati, priklic pada sorazmerno z "rezanjem" števila kandidatov, vendar dosega pri prvih 5.000 kandidatih (kar je manj kot 10 % celotnega seznama izluščenih) še vedno 0,75 (oziroma 0,63), kar je dober rezultat.

4. Sklep

Pri vsakem ročnem prepoznavanju terminov v besedilu naletimo vsaj na štiri težavne točke, ki so: meja med terminološko in splošno leksiko, razmerje med terminologizacijo in determinologizacijo, termini več strok ter terminološke kolokacije. Zato je še toliko bolj pomembno, da imamo objektiven kazalec večje oz. manjše potencialne terminološkosti. Programu, ki deluje na podlagi statističnih izračunov in vnaprej danega, pri tokratnem eksperimentu precej obsežnega nabora jezikoslovnih vzorcev, smo dali prav tako nalogo: da v korpusu besedil odnosov z javnostmi prepozna strokovno izrazje. S terminološko utežjo, katere izhodišče je razmerje relativnih pogostosti besed v specializiranem korpusu in splošnem korpusu – v našem primeru KoRP in FidaPLUS – smo dobili več seznamov; v vrhu enega od njih so npr. samostalniki *javnost*, *odnos*, *organizacija*, *komuniciranje*, *podjetje*, *medij* in *znamka*. Pred njihovo dokončno vključitvijo v geslovnik bo še vedno potrebna analiza področnih strokovnjakov ter sodelujočih jezikoslovcev, a tak pristop kakovostno vendarle močno presega zgolj individualne in večkrat hipne odločitve posameznikov, ki se iskanja terminov lotijo ročno.

Obe primerjavi, tj. primerjava seznama terminološko izluščenih samostalnikov in pogostostnega seznama samostalnikov ter ocena terminološkosti obeh seznamov s strani strokovnjakov odnosov z javnostmi, sta pokazali prednost izluščenega seznama. Razlike so morda na prvi pogled majhne, vendarle pa so zelo relevantne, saj potrjujejo ravno tisto, kar nas je še posebej zanimalo: občutljivost terminološke uteži. V zvezi s slednjo v nadaljevanju raziskave načrtujemo še en preizkus, in sicer preverbo vpliva strokovnega področja na večjo oz. manjšo uspešnost rezultatov luščenja v razmerju do terminološke uteži. Ta hip se namreč zdi, da je družboslovno področje, kakršno so odnosi z javnostmi, ki obstajajo na presečišču menedžmenta, marketinga in komunikologije (Gruban, 1998: 25) zaradi delne tematske prekrivnosti z referenčnimi korpusi (časopisi in revije, deloma tudi knjige z besedili o aktualnih dogodkih, gospodarstvu, poslu, financah ipd.) za luščenje, ki v uteži vključuje ravno tovrstno primerjavo, večji izziv kot katero drugo specializirano področje ali podpodročje, npr. polimerno inženirstvo ali zaključni procesi v biotehnologiji.

Odločitev za luščenje glagolskih vzorcev, sploh v tolikšnem obsegu, se jezikoslovnemu bralcu morda zdi presenetljiva. Presodili smo, da je glagolom kot pogosto spregledanemu delu terminologije (prim. npr. Žele, 2004: 78: glagoli so "prav zaradi svoje organizacijske vloge v stavčnih povedih povsem netipična besedna vrsta za termine") vredno dati tovrstno pozornost ter z veliko količino podatkov in statistično podprto potrditi ali ovreči njihovo terminološko relevantnost. Z dopuščanjem morebitne specifičnosti katere od strok si dovoljujemo posplošitev, da luščenje glagolskih oblikoskladenjskih vzorcev (za razliko od enobesednih glagolov) za slovenščino ne daje relevantnih seznamov z večbesednimi glagolskimi termini (prim. tudi Arhar Holdt, 2011: 121–125). Ob možnosti opazovanja tipičnega besedilnega okolja glagolov s katerim drugim orodjem je luščenje glagolskih vzorcev bolj smiselno povsem opustiti. Po drugi strani pa velja poudariti, da so bili veliko boljši kot pri glagolskih vzorcih rezultati luščenja vzorcev, ki jih tu nismo posebej obravnavali, tj. vzorcev s samostalniškim jedrom. Tako so npr. vzorci *P S*, *P in P S* ter *P P S* v zgornjem delu seznama v več kot 50 % dali gradivo, ki je neposredno uporabno za slovarski geslovnik. Odprto zaenkrat ostaja vprašanje, ali je smiselno še povečati dolžino vzorcev na pet, šest ali več besed. V naši analizi smo se ustavili pri štiribesednih, vendar tej meji nismo pripisovali dokončnosti. Dve petbesedni enoti sta kot termina prepoznala področna strokovnjaka, kar jasno kaže, da v odnosih z javnostmi taki termini so. Ena pot do njih je povečanje obsega vzorcev luščenja, druga pa analiza besedilnega okolja.

Analiza priklica še potrjuje naše prepričanje, da smo pri izboljšavah luščilnika ter pri širjenju seznama vzorcev na pravi poti, saj je samodejno luščenje pri terminih, ki sta jih označila oba strokovnjaka, doseglo kar 93-odstotni priklic. Po drugi strani pa se skozi nizko ujemanje med obema strokovnjakoma jasno kaže subjektivnost same definicije terminološkosti ter posledično zahtevnost zastavljenih ciljev.

Zahvala

Avtorice se zahvaljujejo anonimnima recenzentoma ter uredniku zbornika za koristne pripombe in predloge. Raziskava, predstavljena v prispevku, je nastala v okviru projekta *Terminološke baze podatkov kot osnova strokovnih znanj: model za sistematizacijo terminologij* (http://www.termis.fdv.uni-lj.si/), ki jo po pogodbi št. 1000-11-274193 financira Javna agencija za raziskovalno dejavnost Republike Slovenije ter sofinancerja Pristop, d. o. o., in Gospodarska zbornica Slovenije. Projekt podpirajo tudi sponzorji: Elektro Ljubljana, d. d., Mercator, d. d., Pošta Slovenije, d. o. o., in Zavarovalnica Maribor, d. d.

Literatura

Arhar Holdt, Š., 2011. *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Atkins, B. T. S., Rundell, M., 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Bergenholtz, H., Tarp, S., ur., 1995. *Manual of Specialised Lexicography*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Biber, D., Conrad, S., Reppen, R., 1998: *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Čermák, F., ur., 2011. *Korpusová lingvistika*. Praga: Nakladatelství lidové noviny, Ústav Českého národního korpusu.

Gantar, P., 2007. *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: Založba ZRC, ZRC SAZU.

Gantar, P., 2009. Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo*, 54(3/4): 69–94.

Gorjanc, V., 2005. *Uvod v korpusno jezikoslovje*. Domžale: Založba Izolit.

Gorjanc, V., Krek, S., ur., 2005. *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina.

Gruban, B., 1998. Izobraževanje v odnosih z javnostmi. V B. Gruban, D. Verčič, F. Zavrl (ur.), *Preskok v odnose z javnostmi*: 25–44. Ljubljana: Pristop.

Halliday, M., et al., 2004. *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum.

Hanks, P., 2008. *Lexicology*. New York: Routledge.

Heid, U., 2006. A model for a multifunctional dictionary of collocations. V *EURALEX*: 979–988.

Holozan, Peter (2006): Dodatne dvoumnosti zaradi popustljivosti analizatorja pri analizi slovenskih stavkov. V T. Erjavec, J. Žganec Gros (ur.), *Jezikovne tehnologije*: 146–149. Ljubljana: Institut Jožef Stefan.

Krek, S., 2003. Sodobna dvojezična leksikografija. *Jezik in slovstvo*, XLII(1): 45–60.

Krek, S., Kilgarriff, A., 2006. *Slovene Word Sketches*. Dostopno prek: <http://nl.ijs.si/is-ltc06/proc/12_Krek.pdf> .

Leech, G., 1992. Corpora and Theories of Linguistic Performance. V J. Svartvik (ur.), *Directions in Corpus Linguistics*: 105–122. Berlin: Mouton de Gruyter.

Logar, N., 2007. *Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Logar, N., Vintar, Š., 2008. Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izrazja. *Jezik in slovstvo*, LIII(5): 3–17.

McEnery, T., Wilson, A., 1996. *Corpus Linguistics*. Edinbourgh: Edinbourgh University Press.

Pearson, J., 1998. *Terms in Context*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Romih, M., Holozan, P., 2002. Infrastruktura za razvoj jezikovnih tehnologij – korpus FIDA in sistem ASES. V T. Erjavec, J. Žganec Gros (ur.), *Jezikovne tehnologije*: 166. Ljubljana: Inštitut Jožef Stefan.

Schryver, G. de, 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(1): 143–199.

Sinclair, J., 2004. *Trust the Text: Language, Corpus and Discourse*. London, New York: Routledge.

Teubert, W., Krishnamurty, R., ur., 2007. *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge.

Vintar, Š., 2003. *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Vintar, Š., 2008. *Terminologija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

Vintar, Š., 2009. Samodejno luščenje terminologije – izkušnje in perspektive. V N. Ledinek, M. Žagar Karer, M. Humar (ur.), *Terminologija in sodobna terminografija*: 345–356. Ljubljana: Založba ZRC, ZRC SAZU.

Vintar, Š., 2010. Bilingual Term Recognition Revisited: The Bag-of-Equivalents Term Alignment Approach and its Evaluation. *Terminology*, 16(2): 141–158.

Vintar, Š., Erjavec, T., 2008. iKorpus in luščenje izrazja za Islovar. V T. Erjavec, J. Žganec Gros (ur.), *Jezikovne tehnologije*: 65–69. Ljubljana: Institut Jožef Stefan.