Nataša Logar[1] in Špela Vintar[2]
[1]*Univerza v Ljubljani, Fakulteta za družbene vede*
natasa.logar@fdv.uni-lj.si
[2]*Univerza v Ljubljani, Filozofska fakulteta*
spela.vintar@guest.arnes.si

UDK 811.163.6'374.26:004.42

KORPUSNI PRISTOP K IZDELAVI TERMINOLOŠKIH SLOVARJEV: OD BESEDNIH SEZNAMOV IN KONKORDANC DO SAMODEJNEGA LUŠČENJA IZRAZJA

Prispevek opisuje postopke izbora in obdelave gesel za namen gradnje terminološkega slovarja odnosov z javnostmi. V uvodnem delu je podana utemeljitev korpusnega pristopa h gradnji terminoloških slovarjev, saj korpus ponuja številne tehnološko in metodološko naprednejše možnosti vpogleda v inventar terminov na določenem področju. Osnova za izbor gesel so samodejno izdelani seznami terminoloških kandidatov, se pravi besed in besednih zvez, ki na podlagi statističnih obdelav in oblikoskladenjskih značilnosti izstopajo kot za stroko pomembne. Ročni pregled tako pridobljenih seznamov pokaže njihovo uporabnost, pa tudi nekatere probleme, povezane z razlikovanjem terminov od neterminov in terminov od terminoloških kolokacij.

1 Uvod

Slovenski terminološki slovarji trenutno nastajajo brez korpusov – le redke izjeme, npr. *Islovar* (http://www.islovar.org; Erjavec in Vintar 2004; Puc in Erjavec 2006) in *večjezični vojaški slovar* (Gorjanc in Logar 2007), pravzaprav le pregovorno potrjujejo pravilo. Ker veliko strok pri nas še nima svojega (sodobnega) slovarja, je izid vsakega tovrstnega priročnika – če ta vsaj v grobem ne nasprotuje temeljnim leksikografskim in strokovnopodročnim normam ter spoznanjem – seveda vseeno dobrodošel.

Za prvi slovenski terminološki slovar velja delo pravnika in jezikoslovca Mateja Cigaleta *Znanstvena terminologija s posebnim ozirom na srednja učilišča* (1880). Osrednja slovarska pozornost je bila na Slovenskem v preteklih stoletjih sicer namenjena izdelavi splošnih slovarjev (čeprav je Levstik na potrebo po tem, da bi imele stroke tudi svoje »posebne slovarje« opozoril že leta 1860). Moderna slovenska leksikografija (Vidovič Muha 2000: 11–16) se je začela s snovalci, kritiki in sodelavci Pleteršnikovega slovarja (1894/95) ter nadaljevala s SSKJ (1970–1991). Spoznanja

in izkušnje slovenske leksikografije, ki je v knjige zapisovala slovar slovenščine, so močno vplivala tudi na to, kako je potekalo uslovarjanje specifičnega segmenta tega slovarja: slovenskih terminologij. Do pred kratkim se tako leksikografija ni razvijala kot samostojno raziskovalno področje, temveč je bila razumljena bolj kot uporabni del leksikologije, kar je prineslo močno zakoreninjeno predstavo o tem, da leksikografija temelji na spoznanjih leksikologije in nima svojega metodološkega ter posledično terminološkega aparata; enako je bilo razumljeno tudi razmerje med vedo o terminih in predstavitvijo terminoloških enot v slovarjih.[1] Takó terminologiji kot leksikografiji je sicer skupno, da se – prek terminov in leksemov – ukvarjata z besedami, saj so takó termini kot leksemi sestavljeni iz ene ali več besed, a taka skupna kategorija lahko usmerja pozornost na skupne lastnosti in jo odvrača od tistih lastnosti, ki ju ločijo:

> Izrazijski (= terminološki) slovarji se od drugih ločijo po tem, da obravnavajo strokovno izrazje (têrmine), zaradi česar kažejo nekaj samosvojih značilnosti ne le v izboru gesel, ampak tudi v drugih pogledih; toda teoretična spoznanja, ki veljajo za slovarje na sploh, veljajo tudi zanje. (Gjurin 1986: 151.)

Meyer in Mackintosh (1996) sta terminografijo kot pripravljanje, sestavljanje in pisanje terminoloških slovarjev ločili od leksikografije zaradi več značilnosti, med njimi je najpomembnejša tista, ki nas bo v nadaljevanju prispevka z vidika korpusov najbolj zanimala, opozarja pa nanjo posredno tudi Gjurin: avtorji terminoloških slovarjev (terminografi) morajo za razliko od avtorjev splošnih slovarjev (leksikografov) svojo osnovno slovarsko enoto, tj. termin, šele prepoznati oz. jo najti. Terminografa torej poleg tega, da so termini kot poimenovanja za specializirane pojme določenega strokovnega področja praviloma enopomenski, neekspresivni, ustaljeni, sistemski in dogovorjeni, ne zanima vsa leksika iz besedilnega gradiva, ampak samo določena.[2]

2 Problem

Kljub stavku, ki smo ga kot prvega zapisali v uvod, bo naše nadaljnje razpravljanje izhajalo iz predpostavke, da imamo korpus strokovnih besedil, katerega namen je pridobitev jezikovnih podatkov za terminološki slovar določene stroke in pri gradnji katerega so se skušala predhodno prepoznati in uresničiti merila (i)zbiranja besedil, po katerih je za korpus mogoče reči, da teži k zajetju čim več terminov določenega področja. Zanimala nas bo torej naslednja terminografska faza: kako v korpusu strokovnih besedil prepoznati oz. najti termine; z drugimi besedami: kako in koliko so pri pripravi geslovnika terminološkega slovarja, tj. nabora leksikalnih enot, ki bodo deležne samostojne celovite terminografske obdelave, v pomoč različni samodejno generirani seznami iz korpusa strokovnih besedil. Ti seznami so lahko eno- in večbesedni. V osrednjem delu prispevka se bomo posvetili prvim, v zadnjem delu pa tudi drugim, vključno s premislekom o meji večbesednih terminov oz. o razmerju termin : terminološka kolokacija.

---

[1] Izraz *terminografija* imamo v slovenskem prostoru sicer vsaj že 20 let (Radovanović 1987).

[2] V tem pogledu so v enakem položaju tudi avtorji frazeoloških slovarjev.

3 Gradivo

Gradivo za analizo je korpus besedil odnosov z javnostmi *KoRP*, ki je od julija 2007 brezplačno javno dostopen na <http://www.korp.fdv.uni-lj.si>. Korpus je jezikoslovno označen (Erjavec 2003: 70–71; Erjavec in Vintar 2004: 100; Gorjanc 2005: 56–70), delo z njim pa poteka v Amebisovem konkordančniku ASP32. Vsebuje 1,824.699 pojavnic. Je enojezični, sinhroni, pisni in trenutno statični korpus strokovnih besedil. Z zadostitvijo vnaprej opredeljeni mreži meril za nabor in izbor besedil skuša čim bolje prestavljati celotno stroko odnosov z javnostmi in je prvi tovrstni korpus strokovnih besedil v Sloveniji (več o zgradbi in drugem gl. na korpusovi spletni strani).

Odločitev, da k izdelavi terminološkega slovarja pristopimo na korpusni način, izhaja iz prepričanja, da je zajemanje in opisovanje besedišča s pomočjo računalniških konkordanc neprimerno lažje in hitrejše kot na osnovi listkovnega gradiva. Izpis konkordanc in njihovo urejanje po levih ali desnih kolokatorjih nam pokaže kolokabilno (tj. povezovalno) in frazno (tj. besednozvezno) obnašanje iztočnice, različni sobesedilni vzorci pa nam pomagajo razbrati njene pomene in rabo. Če je korpus oblikoskladenjsko označen, lahko okolje iztočnice raziskujemo še bolj usmerjeno, npr. tako, da si ob samostalniškem geslu prikažemo vse pridevnike ali predložne zveze. Pogostost pojavitve je pri tem pomembno – a nikakor ne edino – merilo pri izdelavi geslovnika in pri opisu določenega gesla.

4 Analiza

Iz korpusov – sploh če so lematizirani in oblikoskladenjsko označeni – je mogoče samodejno pridobiti zelo različne sezname besed in zvez besed (v tem pogledu še dodatne možnosti odpira členjenost na podkorpuse), vsekakor pa je rezultat osnovne analize vsakega korpusa lista besed.

4.1 Lista besed

Spodnja tabela prikazuje listo besed oz. pogostostni seznam lem v korpusu besedil odnosov z javnostmi *KoRP* v primerjavi z enakim seznamom v referenčnih korpusih slovenskega jezika *FIDA*[3] in *FidaPLUS*.[4] Poudarjeno tiskane so tiste leme, ki jih v

[3] Korpus *FIDA* je rezultat projekta dveh pedagoško-raziskovalnih in dveh komercialnih partnerjev: Filozofske fakultete Univerze v Ljubljani, Instituta Jožef Stefan, založbe DZS, d. d., in podjetja Amebis, d. o. o. Projekt sta v celoti financirala oba komercialna partnerja (www.fida.net, 20. 8. 2007). Podatke v tabeli objavljamo po Gorjanc 2005: 73.

[4] Korpus *FidaPLUS* je rezultat aplikativnega raziskovalnega projekta *Jezikovni viri za slovenščino* (L6-5409), ki ga je financirala Javna agencija za raziskovalno dejavnost Republike Slovenije, sofinancirali pa založba DZS, d. d., in podjetje Amebis, d. o. o. Poleg Filozofske fakultete Univerze v Ljubljani kot nosilne ustanove sta pri projektu sodelovali še Fakulteta za družbene vede Univerze v Ljubljani ter Institut Jožef Stefan. Gradnja korpusa se je deloma financirala tudi iz ciljnih raziskovalnih projektov *Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika* (V6-0122) ter *Oblikovanje slovenskega korpusnega omrežja* (V6-0121) (www.fidaplus.net, 20. 8. 2007). Za

korpusu *FidaPLUS* med prvimi 30 lemami ni.

Korpus besedil odnosov z javnostmi *KoRP*	**Korpus slovenskega jezika *FIDA***	**Korpus slovenskega jezika *FidaPLUS***
biti	biti	biti
in	v	v
v	in	in
z	na	na
na	za	se
za	da	z
ki	ta	za
se	ki	da
javnost	pa	on
da	z	ki
odnos	tudi	pa
ta	s	ta
on	po	tudi
pa	kot	ne
organizacija	še	po
tudi	ves	še
ali	iz	kot
ne	ali	ves
kot	o	leto
o	tako	iz
lahko	imeti	o
podjetje	jaz	pri
pri	lahko	imeti
svoj	drug	od
komuniciranje	nov	jaz
ves	morati	do
medij	slovenski	ali
po	prvi	že
med	čas	svoj
kateri	dan	lahko

Tabela 1: *Lista 30 najpogostejših lem v korpusu besedil odnosov z javnostmi* KoRP *in korpusu slovenskega jezika* FIDA *ter korpusu slovenskega jezika* FidaPLUS.

podatke v tabeli se zahvaljujemo koordinatorju projekta Simonu Kreku.

Med 30 najpogostejšimi lemami v korpusu besedil odnosov z javnostmi je 13 lem, ki jih ni med 30 najpogostejšimi v korpusu *FIDA*, in 8 lem (tiskane krepko), ki jih ni med 30 najpogostejšimi v korpusu *FidaPLUS*. Od slovničnopomenskih besed med zadnjih 8 sodita zaimek *kateri* in predlog *med*, od ostalih besed pa je med najpogostejšimi 30 lemami v korpusu besedil odnosov z javnostmi 6 samostalnikov, ki se v ostalih dveh korpusih niso uvrstili tako visoko, in prav vsi ti samostalniki so v strokovni komunikaciji odnosov z javnostmi terminološki: *javnost*, *odnos*, *organizacija*, *podjetje*, *komuniciranje* in *medij*.

Z vidika terminološkosti smo pregledali tudi preostali del liste besed, kot bomo videli, pa je tovrstno ocenjevanje zlasti pri samostalnikih in glagolih subjektivno ter nujno zahteva vsaj še vpogled v besedilno okolje.

Odnosi z javnostmi so kot mlada stroka (za začetek velja leto 1990) delno prekrivni vsaj s trženjem in menedžmentom. Naše izhodišče razumevanja terminov kot terminov (tudi) odnosov z javnostmi (in ne samo terminov neke druge stroke) je bilo zato naslednje: vse besede ali besedne zveze, ki imajo specializirano referenco, je treba ne glede na strokovno področje, ki so mu pripadale kot prvotnemu, takrat ko postanejo del slovarja drugega področja, razumeti tudi kot del terminologije tega drugega strokovnega področja (Pearson 1998: 13, 87). Treba je upoštevati tudi spoznanje, da večja pogostost v korpusih strokovnih besedil ne prinaša nujno tudi večje terminološkosti, kar pomeni, da je treba ozko specializirane termine iskati tudi ali pa predvsem v tistem delu liste besed, kjer je število pojavitev manjše in majhno. Kennedy (1999: 100) je sicer celo za več kot petmilijonski korpus splošnega jezika *American Heritage Intermediate Corpus* (1971) ugotovil, da se v njem skoraj 40 % besed pojavi samo enkrat (*hapax legomena*), kar kaže, da tudi korpus take velikosti ni trdna osnova za leksikografsko proučevanje besed z nizko pogostostjo, vsekakor pa je treba enkratnim pojavitvam prav v korpusih strokovnih besedil posvetiti posebno pozornost, saj se v njih, kot rečeno, »zelo specifična poimenovanja z visoko terminološko vrednostjo /…/ pogosto pojavijo le enkrat« (Vintar 2003a: 69). To pa pomeni, da vsaj okvirna odločitev, kakršno lahko sprejmejo avtorji splošnega slovarja, tj. da bodo iz sto- ali večstomilijonskega referenčnega korpusa kandidate za geslovnik izbrali izmed različnic z npr. vsaj sto pojavitvami,[5] pri korpusnem terminološkem slovarju ni ustrezna, pa čeprav bi bilo število mejnih pojavitev ustrezno nižje.

4.1.1 Lista besed – pridevniki

Pridevniki samostojno sicer niso termini, ker pa za vrstne pridevnike velja, da skupaj s samostalniki tvorijo stalne besedne zveze (Vidovič Muha 2000: 69, 310–328), so

[5] Za primerjavo: vprašanje o pogostostni meji, ki še zadošča za vključitev leme v slovar, pri listah besed iz sto- ali večstomilijonskih korpusov si postavlja Šulc (2002) in pri stomilijonskem referenčnem korpusu predlaga za izhodiščni seznam pred končnim geslovnikom mejo petih pojavitev (prav tam: 218, 219).

kandidati za termine odnosov z javnostmi zveze samostalnikov s pridevniki kot[6] *komunikacijski*, *blagovni*, *javni*, *lokalni*, *poslovni*, *družbeni*, *organizacijski*, *medijski*, *strateški*, *finančni* itd. Lastnostnih pridevnikov je na listi do 1000 pojavitev malo: *velik*, *različen*, *pomemben*, *nov*, *dober* in *določen*. Ker so pravi kakovostni pridevniki, pa tudi del mernih pridevnikov v terminološkem slovarju aktualni le kot kolokatorji, si neposredno z liste besed z njimi v slovarju ne moremo pomagati, vsekakor pa je lista besed dobro izhodišče za prepoznanje večbesednih terminov z vrstnopridevniškimi prilastki.

4.1.2 Lista besed – glagoli

Terminologija kot nosilec pojmovnega sveta stroke se običajno veže na samostalniško leksiko, prim. npr. prepoznanje večinskosti samostalniške terminološke leksike pri Vidovič Muha (2000: 117) oz. povzetek Žele (2004: 78), da so glagoli »prav zaradi svoje organizacijske vloge v stavčnih povedih povsem netipična besedna vrsta za termine«, ob opombi, da je v češki literaturi podan podatek o 7,2 % glagolskih terminov (nasproti 92,4 % samostalniških) (Žele, prav tam). Pričakovana samostalniškost terminov, vezana na pojem kot statično predmetnost, je pogosto razlog, da neupravičeno spregledamo glagolske termine (pa tudi termine drugih besednih vrst, npr. prislovne *allegro*, *forte*, *rubato* itd. v glasbi). Kot glagolski izjemi v sicer večinsko samostalniški terminološkosti se večkrat navajata vojaška in športna terminologija. V SSKJ so npr. s terminološkim kvalifikatorjem vojaško označeni glagoli *blindirati*, *degazirati*, *demaskirati* itd., s kvalifikatorjem športno pa *centrirati*, *deskati*, *diskvalificirati* itd. Tudi npr. ni dvoma, da naslednje pojasnilo prepoznava glagole *dopolniti*, *izpopolniti*, *popolniti* in *nadomestiti* kot vojaškoterminološke enote:

> Postavlja se vprašanje, ali je v vojaškem izrazju med /naštetimi/ glagoli mogoče (in treba) pomensko razlikovati ali ne. /…/ Razlog, da se je spričo Vojaškega slovarja vprašanje /…/ pojavilo, je izrazoslovnega značaja, kar pomeni, da vsaka stroka želi in mora videti zadevno resničnost natančneje od splošnega, nestrokovnega videnja in to resničnost seveda tudi natančneje poimenovati. Za to nalogo lahko uporabi besede, med katerimi se v strokovni sferi ne ločuje, izrazoslovna praksa pa velikokrat (in prav uspešno) vrsto sinonimnih besed uporabi v stroki tako, da eno besedo uporablja zmeraj v tej, drugo pa v oni besedni zvezi. / Take, strokovno utemeljene pomenske razlike morajo registrirati terminološki slovarji. (Korošec 1998: 88.)

V zvezi z glagoli kot vojaškimi termini sodi sem kot potrditev vsaj še prispevek *O ločevanju med pomenoma glagolov* streljati *in* obstreljevati (Korošec 1998: 125–127), v katerem uvodoma piše:

> Vsaka stroka teži h kar se da natančnemu sporočanju o strokovnih vsebinah in to seveda velja tako za poimenovanja predmetnega in pojmovnega sveta, torej na ravni besed – terminov – kakor tudi na ravni izjav o svetu stroke, se pravi na ravni stavkov. / Za natančno ločevanje med pojmi morajo biti na razpolago različne besede, in ker v vojaški stroki

[6] Vse tovrstne ocene v prispevku so zgolj jezikoslovne in so podane brez posveta s področnim strokovnjakom.

> obstaja potreba po ločevanju med pomenoma, katerih vsebina sta različni dejavnosti, ena splošnejša, druga specialnejša, je treba za to uporabiti različne besede.

Glagolska gesla imata npr. tudi *Islovar* (2001) in *Planinski terminološki slovar* (2002). Kot samoumevno dejstvo so glagoli med matematične termine prišteti tudi v Gorjanc 1995/96, saj avtor zanje ugotavlja (670): »Med glagolskimi terminološkimi besednimi zvezami so matematične vezavne s tožilnikom: *eliminirati neznanko*, *krajšati ulomek*.« Enako je v zvezi z glagoli kot termini razmišljanje Erjavec in Vintar (2004: 104 – gl. v nadaljevanju; prim. tudi Vintar 2008: 40–41).

Kot kandidati za termine odnosov z javnostmi se na listi besed kažejo glagoli *sporočiti*, *komunicirati*, *objaviti*, *skupiniti*, *prikrojiti*, *akreditirati*, *fokusirati*, *lansirati*, *dezinformirati*, *mrežiti* itd. Vendar je za razpoznanje glagolov, ki so v odnosih z javnostmi terminološki, zgolj ogled seznama premalo, saj so ob naštetih terminološko predvidljivejših primerih na listi besed npr. še *poslovati*, *vlagati*, *svetovati* in *upravljati* ter *ugotoviti*, *upoštevati*, *temeljiti* in *obravnavati* – za prve bi morda lahko rekli, da so vendarle značilni za odnose z javnostmi, drugi pa so del splošnega slovarja jezika in kot taki pač značilen del tudi strokovnih besedil.[7]

4.1.3 Lista besed – samostalniki

Tudi ločevanje samostalnikov na tiste, ki so zelo verjetno termini, in na tiste, ki so del splošnega jezika, in to zgolj na podlagi seznama, se je izkazalo za vsaj deloma subjektivno. Na prvi pogled se je zdelo, da med najpogostejšimi samostalniki v korpusu med terminološke kandidate sodijo: *javnost*, *organizacija*, *podjetje*, *komuniciranje*, *medij*, *znamka*, *novinar*, *upravljanje*, *komunikacija*, *deležnik*, *oglaševanje*, *trg*, *marketing*, *management*, *storitev*, *potrošnik*, *kampanja* itd., potem pa smo začeli razvrščati primere v drugo (splošnojezikovno) skupino in ugotovili, da npr. *potrošnik*, *podjetje* in *novinar*, ki smo jih uvrstili med terminološke kandidate, po naši presoji vendarle niso nič manj del splošnega jezika kot *informacija*, *okolje*, *sporočilo*, *ugled*, *razvoj* itd., ki so se nam prvotno bolj zdeli del splošnega jezika. Terminološko očitno prepoznaven je bil na ta način le samostalnik *deležnik*.

Zadrego, v kateri smo se znašli pri samostalnikih, že prej pa tudi pri glagolih, opisuje že Pearson (1998: 26–28). Avtorica izhaja iz spoznanja, da je način, na katerega ljudje govorijo o stvareh, odvisen od konkteksta, v katerem so, in od védenja, ki ga imajo. Avtorica meni (27), da ostaja soodvisnost med številom ljudi, ki poznajo leksiko določenega specializiranega področja, in percepcijo te leksike kot specializirane: manj, kot je takih ljudi, bolj je verjetno, da bo taka leksika (in področje) dojeta kot specializirana. Kar torej ljudi po občutku usmerja v odgovor na vprašanje, ali je nek leksem termin ali ne, je njegova relativna nepogostost v splošnem jeziku (in

[7] Prim. tudi izhodišče obravnave glagolske terminologizacije pri Žele (2004: 79): »Merilo stopnje terminološkosti določenega glagola je konkretno obravnavano besedilo.«

torej nerazumljivost pomena) in/ali sporazumevalno okolje, v katerem je rabljen. Pearson tudi ugotavlja, da pogostost v splošnem jeziku ne more biti merodajna za opredeljevanje, katera beseda je termin in katera ni, da pa je sporazumevalno okolje (vključno s sobesedilom), v katerem se pojavlja, pri odgovoru na to vprašanje vsekakor treba upoštevati – še več, avtorica verjame (35), da je ustrezna opredelitev sporazumevalnega okolja, za katerega je verjetno, da se bo v njem pojavil termin, najboljša pot do ločevanja med terminološko in neterminološko leksiko. Iz navedenega tudi v kontekstu pričujoče raziskave izpeljujemo spoznanje, da če je strokovno izrazje lastnost strokovnih jezikov in ga lahko opredelimo le na podlagi sobesedila, je edina sprejemljiva metoda za delo korpusni pristop.

S postopki avtomatskega pridobivanja terminoloških kandidatov iz korpusov se ta možnost že uresničuje na raziskovalni, pa tudi komercialni ravni, saj sodobna prevajalska orodja, kot je *SDL Trados* (http://www.trados.com), ponujajo komponente za samodejno statistično luščenje terminov iz besedil.

4.2 Samodejno luščenje terminoloških kandidatov

Kompleksnost zastavljene naloge dobro povzame Sager (1998/99), ko pravi, da so termini pravzaprav besede s specifično funkcijo, ali drugače rečeno (in kot smo nakazali že zgoraj), termini se formalno z ničemer ne ločijo od besed. Če jih želimo samodejno izluščiti iz strokovnih besedil, moramo za to seveda vseeno oblikovati določena – formalna, saj drugačnih še nismo sposobni računalniško obdelovati – merila, ta pa so vselej le grob približek dejanskim značilnostim terminološkega inventarja v korpusu.

Pri luščenju terminov smo uporabili metode, opisane v Justeson in Katz 1995, Jacquemin 2001 ter Vintar 2003b. Iz korpusa besedil odnosov z javnostmi smo poskusno pripravili dva seznama izluščenih terminoloških kandidatov: enobesednega in večbesednega.

4.2.1 Seznam enobesednih luščenih terminoloških kandidatov

Ta seznam je nastal po metodi relativne pogostosti, ki temelji na predpostavki, da se bodo v strokovnih besedilih izrazi, ki jih v splošnem jeziku ni, in izrazi, ki imajo glede na splošni jezik zožen pomen, pojavili relativno pogosteje kot v drugih, nestrokovnih besedilih (Damerau 1993). Razmerje relativnih pogostosti v nadaljevanju imenujemo terminološka utež.

Primerjali smo torej relativne pogostosti liste besed korpusa *KoRP* in liste besed korpusa *FIDA*. Nastal je seznam z 32.032 enobesednimi enotami, urejen padajoče po terminološki uteži. Natančneje smo si ga za ta prispevek ogledali na dveh mestih:
a) prvih 1000 enot, ki imajo vrednost nad 100, kar pomeni, da gre za besede, ki jih v korpusu *FIDA* ni, in

b) prvih 1000 enot, ki imajo vrednost pod 100, kar pomeni, da v korpusu *FIDA* so (v seznamu zasedajo mesta od 11.124 do 12.124).

K **a)**: Pregled prvih 1000 enot na seznamu je pokazal, da so razen enot *deležniški* in *deležnik*, ki sta očiten del termina oz. termin (na seznamu zasedata 1. in 2. mesto), preostale enote predvsem:
- lastna imena,
- zatipkane besede,
- številke,
- angleške besede (predvsem kot del bibliografskih enot),[8]
- redke besede in neologizmi, vendar na prvi pogled ne terminološki ali vsaj ne terminološki v odnosih z javnostmi, npr. *spihovati*, *paternalizem*, *nudenje*, *fenski*, *biltenov*, *varietejski*, *etabliran*, *razminiranje*, *publicirati*, *nepremoženjskost*, *preokupiranost*.

Termini so v tem delu seznama dokaj skriti, čeprav seveda so tu. Domnevamo lahko, da je izrazje odnosov z javnostmi (verjetno to velja za več družboslovnih ved in tudi širše) v veliki meri tudi del splošnega jezika, tako da zgolj primerjava (ne)pojavitve v splošnem korpusu prinese veliko odvečnega gradiva. Dodatna proučitev bi pokazala tudi, ali gre v tem smislu za z vidika luščenja »neugoden« korpus, »ki sicer zajem/a/ besedila določene stroke /…/ in /je/ torej besedilnovrstno homogen, posamezna besedila pa vendarle vsebujejo izrazje z zelo različnih področij« (Vintar 2002: 80), kar daje slabše rezultate pri avtomatskem pridobivanju terminoloških kandidatov.

K **b)**: Pregled prvih 1000 enot na seznamu, ki v korpusu *FIDA* so, je pokazal, da je lastnih imen tu veliko manj (približno 2 %), da je približno 30 % enot številk in da je zatipkanih besed tu malo. Vse drugo pa je že vredno nadaljnje proučitve. Na tem mestu se bomo na kratko ustavili le pri glagolih.

Glagolov je v tem delu okrog 60. Zgoraj smo videli, da bo terminološke glagole težko prepoznati, čeprav v slovar sodijo. V tem delu seznama so glagoli: *soudeležiti*, *ponderirati*, *zainteresirati*, *strojiti*, *okvirjati*, *komunicirati*, *pristopati*, *opredmetiti*, *sporočiti*, *implementirati*, *zapozneti*, *duhati*, *indeksirati*, *ugledati*, *procesirati*, *ciljati*, *podlagati*, *razpotegovati*, *optimizirati*, *operacionalizirati*, *zajemati*, *decentralizirati*, *vrednotiti*, *letati*, *informirati*, *priporočiti*, *odločevati* itd. Del teh glagolov je navrgel korpusni šum, drugi so – kot smo že ugotavljali – verjetno značilni za strokovna besedila nasploh (kar bi lahko statistično potrdila ali ovrgla primerjava s korpusom strokovnih besedil drugega področja), tretji pa so značilni za odnose z javnostmi: *komunicirati*, *okvirjati*, *sporočiti*, *ciljati*, *informirati*, *odločevati*. Seveda bi bil nujen še pregled besedilnega okolja, a seznam razmeroma dobro kaže, katero izrazje si je vredno ogledati še podrobneje.

[8] Ker se med njimi lahko skriva termin, ki v slovenščini še nima svojega ustreznika, jih bo področni strokovnjak vendarle moral pregledati.

4.2.2 Seznam večbesednih luščenih terminoloških kandidatov

Ta seznam je nastal s kombinacijo statistične in jezikoslovne metode. Iskali smo naslednje skladenjske vzorce:

Prid + Sam
Sam + Sam
Sam + Sam + Sam
Prid + Sam + Sam
Sam + Prid + Sam
Sam + Predl + Sam
Sam + Predl + Sam + Sam
Sam + Predl + Prid + Sam,

ki jim je bila dodeljena terminološka utež, sestavljena iz normaliziranega seštevka terminoloških uteži posameznih sestavin besedne zveze.

Pred ogledom večbesednega seznama luščenih terminov je treba opozoriti še na pojav *terminoloških kolokacij*. Da je enota geslovnika za terminološki slovar *termin*, je definicijska lastnost te vrste slovarjev.[9] Ker pa seznami večbesednih izluščenih terminoloških kandidatov temeljijo na statistično izračunani vrednosti moči medbesedne povezovalnosti, je treba opozoriti, da so na teh seznamih enote dveh vrst: na eni strani (a) poimenovanja, ki zadoščajo sistemskim »zahtevam« po tem, da so kot celota termin (o sistemskih lastnostih strokovnih besednih zvez gl. Vidovič Muha 1988; 2000: 68–70); po drugi strani (b) pa leksikalno in/ali pragmatično povezane ponovljive sopojavitve vsaj dveh leksikalnih enot, ki sta med seboj v neposrednem skladenjskem razmerju (Bartsch 2004,[10] nav. po Heid 2006: 980), se pravi: kolokacije. Povedna sta v tem smislu tudi naslednja dva navedka:

> Študenti so imeli precej težav pri razlikovanju med terminološkimi kolokacijami /.../ in pravimi termini. Tako se na primer pojavi izraz *language technology application*, ki je verjetno kompozitivna kolokacija, kjer se pomen sestavi iz *language technology* in *application*. Čeprav smo v začetku izhajali iz načela, da bomo nediskriminatorno med gesla uvrščali tudi glagolsko izrazje in druge nesamostalniške zveze, se kmalu pokaže, da se nam glagoli kljub pogostosti in specifičnemu pomenu mnogokrat ne zdijo primerni za uvrstitev med iztočnice. Vsaj tisti, ki najbolj odstopajo od svojega splošnojezikovnega pomena, na primer shraniti, brskati /v informatiki/, bi si zagotovo zaslužili terminološko obdelavo. (Erjavec, Vintar 2004: 104.)

[9] Nadaljnje slovarske uresničitve geslovnikov so potem lahko različne: v *Slovenskem elektrotehniškem slovarju* (1957–2001), *Meteorološkem terminološkem slovarju* (1990), *Islovarju* (2001–) itd. je vsaka poimenovalna enota (termin, tudi večbesedni) samostojno geslo, medtem ko so npr. v *Splošnem tehniškem slovarju* (1962/64), *Vojaškem slovarju* (1977/2002) in *Pravnem terminološkem slovarju: do 1990, gradivo* (1999) gesla le enobesedna, torej so kot gesla lahko prikazani le deli poimenovalnih enot ali njihovega besedilnega okolja, npr. v *Vojaškem slovarju* predlogi *čez*, *na*, *ob*, pridevniki *občuten*, *obkoljen*, *kalibrski* in glagoli *iti*, *delati*, *odpreti.*

[10] Bartsch, S., 2004: *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic contraints of lexical co-ocurence*. Tübingen: Narr.

> Čeprav bi pojma /termin in terminološka kolokacija/ v grobem lahko razmejili tako, da za termine štejemo tiste besedne zveze, ki v pojmovnem sistemu področja poimenujejo opredeljene pojme in se zato kot fiksna poimenovanja ne spreminjajo, kolokacije pa so ustaljene jezikovne povezave med termini oziroma med terminom in neterminološkim jezikovnim sredstvom, se v praksi tako posplošena razmejitev ne obnese. /…/ /T/erminološkost besedne zveze /je/ izrazito subjektiven pojem, ki je močno odvisen od uporabnika terminologije. Kar je za terminologa ali dokumentalista zgolj kolokacija, je za prevajalca ali tehničnega pisca termin prav na podlagi kriterija, da gre za edini ustaljeni in sprejemljivi način opisa določenega strokovnega dejstva. (Vintar 2003a: 74.)

Teubert (2005/1999: 106) ocenjuje, da je predvsem področje med leksiko in skladnjo tisto, h kateremu lahko prispeva oz. je prispevalo korpusno jezikoslovje, in meni, da lahko korpusno jezikoslovje s tem, ko daje statistične podatke o značilnem sopojavljanju, veliko bolje kot »klasično« jezikoslovje sledi pojavu posebne pomenske kohezije med kolokacijskimi elementi (113–114). Za razliko od strukturalnega jezikoslovja, za katerega Gantar (2004: 116) ugotavlja, da pri proučevanju kolokacij ni dalo dokončnih odgovorov, ker gre za pojav z nejasno določenimi mejami znotraj strukturalno jasno določenega jezikovnega sistema, se je besedilni pristop (Firth, Halliday, Sinclair idr.) osredotočil na besedno povezovalnost, v kateri leksikalnih enot ni več smiselno razumeti kot dokončnih v smislu razmejitve med besedami in zvezami na eni strani ter med kolokacijami in stalnimi besednimi zvezami na drugi. Vsekakor je proučevanje kolokacij ena od nujnih analiz vsakega korpusa, katerega namen je (terminološki) slovar. Naloga terminografa je zadovoljiti čim več poizvedb različnih uporabnikov slovarja, to pa pomeni, da je treba termine vključiti tako, da bosta celovito terminološko informacijo dobila dokumentalist in prevajalec, pa tudi vsi drugi. V slovarju je torej treba prikazati tako termine kot terminološke kolokacije, zato je pomembno, da so na seznamih iz korpusov, ki jih bo nadalje proučil terminograf skupaj s področnim strokovnjakom, oboji. Vprašanje, ki ga je treba rešiti v nadaljevanju, je način umestitve oz. prikaza enih in drugih v geselskem članku.

Če se torej s to zavestjo vrnemo k pregledu spiska samodejno izluščenih besednih zvez, lahko zapišemo naslednja opažanja. Natančneje smo pregledali prvih 1000 enot: v drugi polovici seznama sicer narašča delež kolokacij, vendar je v tem seznamu veliko terminov. Ko smo izločili lastna imena, je bilo skoraj vse ostalo prepoznavno bodisi kot termin, npr. *blagovna znamka*, *neprofitna organizacija*, *korporativna identiteta*, *komunikacijski menedžment*, *marketinško komuniciranje*, *uglednostni kapital*, *ciljna javnost*, *mnenjski voditelj*, bodisi kot kolokacija, npr. *vloga managerja*, *oddelek za odnose z javnostmi*, *deležnik organizacije*, *zadovoljstvo zaposlenih*, *pozitivna publiciteta*. Malo je bilo tu terminov z drugih področij (vsaj po naši oceni), npr. *kreditna točka*, *postmoderna družba*, *kognitivna disonanca*, *pomenotvorni proces*, ali pa ne zgolj terminov odnosov z javnostmi ali sorodnih področij, npr. *sinergijski učinek*, *diplomska naloga*, *ključna beseda*, *zdravstvena organizacija*, *turistična destinacija*.

Luščenje terminoloških kandidatov je bilo tako po količini in kakovosti razmeroma uspešno. Vseskozi pa se moramo zavedati, da s pregledi takih seznamov ocenjujemo zgolj natančnost metode luščenja, ne pa tudi priklica, ali z drugimi besedami: le iz

pregleda seznama ne bomo nikoli vedeli, koliko večbesednih terminov je ostalo neizluščenih. Vsako resno terminografsko delo vključuje tudi ročno pregledovanje korpusnega gradiva, pa vendar samodejno pridobljeni seznami prinašajo relevantno dopolnitev in kvalitativni presežek drugih metod.

5 Sklep

V prispevku smo predstavili del terminografskega projekta gradnje korpusa in slovarja odnosov z javnostmi, pri čemer smo se še posebej posvetili utemeljitvi korpusnega pristopa, ki po našem prepričanju prinaša tehnološko sodobnejše, metodološko naprednejše in nenazadnje uporabniku prijaznejše rezultate. Obenem ta pristop omogoča delno avtomatizacijo postopka identifikacije terminoloških kandidatov. Kot po eni strani našteto terminografu lajša delo pri opisu segmenta jezika, ga po drugi strani naredi tudi veliko bolj kompleksnega. Zadrega, ali gre za termin ali ne in kje se ta konča, je pri ročnem izpisovanju brez sobesedila bolj ali manj razrešena s hipno odločitvijo izpisovalca, še manjša je, če se geslovnik izdeluje na podlagi že obstoječih slovarjev, pri množici med seboj prepletajočih se korpusnih podatkov pa je mnogo več sivih polj. Korpusi strokovnih besedil z več sto tisoč ali celo milijoni besed so razrahljali več mej: mejo med terminološko in neterminološko leksiko, mejo, do katere še govorimo o večbesedni poimenovalni enoti in čez katero je že prostor kolokacij, ter mejo, ki določa, ali gre za termin področja, ki ga obravnavamo, ali ne. Korpus s frekvenco razvidno izpostavi tudi terminološke kandidate, ki so sistemsko termini, zelo verjetno pa jih nobeno področje ne bo vzelo za svojega, ker so del nekakšnega skupnega slovenskega strokovnega jezika. Zaradi večje količine in boljše kakovosti podatkov se zdi, da je korpusni terminograf še pogosteje prisiljen v tehtanje, zaradi katerega si bo ob konkretnih primerih – paradoksalno – največkrat želel, da bi imel podatkov še več. Metaforično bi lahko rekli, da se v bolj zgoščeni mreži točk tisti, ki te točke povezuje v slovar – in je ob tem odgovorno zavezan nalogi, da dela jezikovni opis –, še bolj zaveda, kaj vse je moral izpustiti in kaj vse je moral posplošiti; dokler namreč teh točk ni poznal, ga niso vznemirjale.

Med vprašanji torej, ki ostajajo odprta in se bo po vsej verjetnosti z njimi treba ukvarjati pri vsakem novem terminološkem projektu, so sestava korpusa strokovnih besedil, merila terminološkosti, obravnava terminoloških variacij in terminoloških kolokacij ter izboljšave metod samodejnega luščenja. Kar se tiče slednjih, smo že v okviru opisanega eksperimenta ugotovili nekatere pomanjkljivosti, ki bi jih bilo mogoče odpraviti z dopolnitvijo seznama luščenih besednovrstnih vzorcev, izločanjem lastnih imen in ciljnim luščenjem nekaterih znanih oblik variacij, denimo kombinacije razvezane oblike termina in njegove kratice. Pomemben pomislek ob takšnih projektih je tudi, da tehnološka infrastruktura, potrebna za izdelavo korpusa, njegovo jezikoslovno označevanje in statistično obdelavo, ni na razpolago vsem, ki se v našem prostoru lotevajo terminografskih projektov. Načrtujemo, da bo omenjeno vrzel v kratkem zapolnil projekt *Slovenski terminološki portal*.[11]

[11] Projekt s šifro L6-9778-0581-06 sofinancira Javna agencija za raziskovalno dejavnost Republike

Viri in literatura

Beran, Jaromir, idr. (ur.), 1999: *Pravni terminološki slovar: do 1990, gradivo*. Ljubljana: ZRC SAZU.

Damerau, Fred J., 1993: Generating and evaluating domain-oriented multi-word terms from texts. *Information processing and management* 29. 433–447.

Erjavec, Tomaž, 2003: Označevanje korpusov. *Jezik in slovstvo* 48/3–4. 61–76.

Erjavec, Tomaž, in Vintar, Špela, 2004: Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika. *Uporabna informatika* 12/2. 97–106.

Gantar, Polona, 2004: *Frazem in njegovo besedilno okolje.* Doktorska disertacija. Mentorica Ada Vidovič Muha. Ljubljana: Filozofska fakulteta.

Gjurin, Velemir, 1986: K začetkom slovenskega slovaropisja. *Slavistična revija* 34/4. 365–392.

Gorjanc, Vojko, 1995/96: Primerjalna razčlenitev terminologije v matematiki in filozofiji. *Jezik in slovstvo* 41/5. 267–276.

Gorjanc, Vojko, 2002: *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov.* Doktorska disertacija. Mentorica: Ada Vidovič Muha. Ljubljana: Filozofska fakulteta.

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

Gorjanc, Vojko, in Logar, Nataša, 2007: Od splošnih do specializiranih korpusov – načela gradnje glede na njihov namen. Orel, Irena (ur.): *Obdobja, metode in zvrsti 24: Razvoj slovenskega strokovnega jezika*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete.

Heid, Ulrich, 2006: A Model for a multifunctional dictionary of collocations. *EURALEX*. 979–988.

Humar, Marjeta, idr. (ur.), 2002: *Planinski terminološki slovar: Slovensko-angleško-nemško-francosko-italijanski slovar planinskega, alpinističnega, plezalskega izrazja*. Ljubljana: Založba ZRC, ZRC SAZU.

Islovar: <http://www.islovar.org>. (Dostopno: november 2007.)

Jacquemin, Cristian, 2001: *Spotting and discovering terms through natural language processing*. MIT Press.

Slovenije, vodi ga Vojko Gorjanc, poleg vodilne Filozofske fakultete Univerze v Ljubljani pa v njem sodelujejo še Institut Jožef Stefan, Fakulteta za družbene vede Univerze v Ljubljani in Amebis, d. o. o.

Justeson, John S., in Katz, Slava J., 1995: Technical terminology: some linguistic properties and an algorihtm for identification in text. *Natural language engineering* 1. 9–27.

Kennedy, Graeme, 1999: *An introduction to corpus linguistics*. London, New York: Longman.

Korpus slovenskega jezika FIDA, 1997–2000, <www.fida.net>. (Dostopno 20. 8. 2007.)

Korpus slovenskega jezika FidaPLUS, 2007, <www.fidaplus.net>. (Dostopno 20. 8. 2007.)

Korošec, Tomo idr. (ur.), 1977/2002: *Vojaški slovar*. Ljubljana: Ministrstvo za obrambo RS.

Korošec, Tomo, 1998: *Slovenski vojaški jezik*. Ljubljana: Fakulteta za družbene vede.

Meyer, Ingrid, in Mackintosh, Kristen, 1996: The corpus from a terminographer's viewpoint. *International journal of corpus linguistics* I/2. 257–285.

Mlakar, France idr., in Ogorelec, Anton idr. (ur.), 1957–: *Slovenski elektrotehniški slovar*. Ljubljana: Elektrotehniška zveza Slovenije; Sloko CIGRÉ.

Pearson, Jennifer, 1998: *Terms in context*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Petkovšek, Zdravko, in Leder, Zvonka (ur.), 1990: *Meteorološki terminološki slovar.* Ljubljana: SAZU, Društvo meteorologov Slovenije.

Pleteršnik, Maks, 1894/95: *Slovensko-nemški slovar.* Ljubljana: Knezoškofijstvo.

Puc, Katarina in Erjavec, Tomaž, 2006: Uporaba korpusa pri urejanju spletnega terminološkega slovarja. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Language technologies/Jezikovne tehnologije IS-LTC.* Ljubljana: Institut Jožef Stefan. 156–161.

Radovanović, Miroslav, 1987: Terminologija – terminografija – tvorba terminov. *Teorija in praksa* 24/10–11. 1453–1462.

Sager, Juan C., 1998/99: In search of a foundation: Towards the theory of the term. *Terminology* 5/1. 41–57.

Slovar slovenskega knjižnega jezika (1970–1991). Ljubljana: DZS.

Splošni tehniški slovar, 1962/1964. Ljubljana: Zveza tehnikov in inženirjev LR Slovenije.

Teubert, Wolfgang, 2005/1999: Korpusno jezikoslovje in leksikografija. Gorjanc, Vojko, in Krek, Simon (ur.): *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina. 103–136./Korpuslinguistik und Lexikographie. *Deutsche Sprache* 4.

Šulc, Michal, 2002: Corpus frequency and lexicographical relevancy: Czech words with a morfem micro- (in hundred million corpus of Czech language – SYN2000). *EURALEX*. 209–220.

Vidovič Muha, Ada, 1988: Nekatere jezikovnosistemske lastnosti strokovnih besednih zvez. Pogorelec, Breda, Sajovic, Tomaž, in Počaj-Rus, Darinka (ur.): *XXIV. seminar slovenskega jezika, literature in kulture*. *Zbornik predavanj*. Ljubljana: Oddelek za slovanske jezike in književnosti Filozofske fakultete. 83–91.

Vidovič Muha, Ada, 2000: *Slovensko leksikalno pomenoslovje. Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Vintar, Špela, 1999: Računalniško podprto iskanje terminologije v slovensko-angleškem vzporednem korpusu. Kovačič, Irena, in Štrukelj, Inka (ur.): *Uporabno jezikoslovje* 7–8. 156–169.

Vintar, Špela, 2002: Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 78–85.

Vintar, Špela, 2003a: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov.* Doktorska disertacija. Mentor: Rastislav Šuštaršič. Ljubljana: Filozofska fakulteta.

Vintar, Špela, 2003b: Kaj izvira iz jezikovnih virov. *Jezik in slovstvo* 48/3–4. 77–88.

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

Žele, Andreja, 2004: Stopnje terminologizacije v leksiki (na primerih glagolov). Humar, Marjeta (ur.): *Terminologija v času globalizacije*. Ljubljana: Založba ZRC, ZRC SAZU. 77–91.