Luščenje specializiranih izrazov za splošni slovar

Špela Vintar in Nataša Logar

Abstract
This paper describes an experiment aimed at extracting specialised vocabulary from specialised subcorpora for the purposes of general lexicography. The main objective was to test the methodology of automatic term extraction, which had been developed for specialised lexicography, in order to gain better insight into the weakness of, and the adjustments required to, the corpus in terms of domain representativeness. The LUIZ Term Extractor was used on two subcorpora, one containing a selection of texts pertaining to physics, biology and chemistry from the ccGigafida corpus, and the other a more homogeneous specialised corpus of textbooks on music theory. The results show that the domain of natural sciences, as represented in the Gigafida corpus, contains few lexical items which require special attention on account of their termhood, whereas a more specialised corpus, as expected, yields a larger number of highly specialised units.

Keywords: term extraction, specialised vocabulary, termhood, general language dictionary

Ključne besede: luščenje izrazja, specializirana leksika, terminološkost, splošni slovar

1 UVOD

Namen pilotnega eksperimenta, katerega rezultate prikazujemo v prispevku, je bil ugotoviti, ali so obstoječe metode luščenja specializiranega izrazja uporabne tudi za potrebe splošnega slovarja oz. kakšne prilagoditve zahtevajo. Dosedanje izkušnje z luščenjem (prim. razdelek 2.1) so bile namreč brez izjeme usmerjene v pridobivanje srednje- in visokospecializiranih izrazov za terminografske namene ali namene modeliranja strokovnih področij, tukajšnja naloga pa je precej drugačna, deloma celo povsem nasprotna.

Z eksperimentom smo želeli preliminarno odgovoriti na tri vprašanja, od tega sta bili dve povezani z gradivom:

1. Koliko gradiva s terminološkim potencialom pokaže Gigafida, če nanjo apliciramo metodo luščenja terminoloških kandidatov, ki je uspešna v specializiranih korpusih strokovnih besedil, na splošnih korpusih slovenščine pa še ni bila preizkušena?
2. Na kolikšen del Gigafide je metodo luščenja terminoloških kandidatov smiselno aplicirati?

Tretje vprašanje je bilo povezano s slovarjem, natančneje z uporabnostjo seznamov izluščenih terminoloških kandidatov pri razdelitvi terminološke (oz. neterminološke) leksike v tri košarice, ki smo jih predvideli za splošni slovar (prim. poglavje Specializirana leksika v splošnem slovarju):

Splošna košarica: Leksika, pri kateri sicer prepoznamo navezavo na določeno strokovno področje, vendar poznavanje področja ni pogoj za njeno razumevanje in pravilno rabo. Pri leksikografskem opisu ne potrebuje področne oznake, njena specializiranost se ne izraža niti v razlagi (*koncert, rastlina, ribič*).

Šolska košarica: Leksika, ki jo redkeje srečamo v splošnih besedilih, a poimenuje temeljne pojme strok in se kot taka pojavlja že v učbenikih nižje stopnje. Njena razlaga lahko vsebuje navedbo področja, zlasti takrat, kadar se determinologizirani pomen razlikuje od področnega (*lestvica, beljakovina, molekula*).[1]

Strokovna košarica: Leksika, ki je splošni javnosti manj znana in je za njeno razumevanje potrebno predznanje. Redko se rabi v splošnih besedilih in se ne determinologizira. Pri njenem leksikografskem opisu se uporabi področna oznaka, ki uporabnika opozarja na terminološkost, razlaga pa je strokovna (*alikvotni ton, hipertrofija, nazivna moč*).[2]

1 Ker so v šolski košarici srednjespecializirani izrazi, se predvideva, da v večini primerov razlago zanje lahko oblikuje leksikograf, pri čemer so v pomoč tudi učbeniška besedila, ki že sama vsebujejo veliko razlag in leksikografu služijo kot predloga.

2 S strokovno razlago mislimo na tip razlage, ki je oblikovno in konceptualno blizu terminološki definiciji in je strokovno pravilna, še vedno pa je prilagojena splošnemu uporabniku in je manj podrobna kot v terminološkem slovarju.

Tretje vprašanje se je torej glasilo: Kako lahko metoda luščenja terminoloških kandidatov leksikografom pomaga pri odločitvah o umestitvi leksike v posamezne kategorije specializiranosti splošnega slovarja?

2 METODA IN GRADIVO

2.1 Luščenje terminoloških kandidatov

Za luščenje smo uporabili luščilnik LUIZ (Vintar 2010), ki iz specializiranega korpusa izlušči terminološke kandidate na podlagi oblikoskladenjskih vzorcev, te pa nato razvrsti glede na t. i. terminološko utež. Slednja je hevristika, ki upošteva ključnost (Scott 1998) posameznih delov terminološke besedne zveze, njeno dolžino ter pogostost v specializiranem podkorpusu.

V dosedanjih raziskavah (Vintar in Erjavec 2008; Logar Berginc in Vintar 2008; Vintar in Fišer 2009; Logar et al. 2012) smo pri luščenju uporabljali obsežen seznam potencialnih oblikoskladenjskih vzorcev za slovenščino, ki se sicer lahko prilagaja specifikam posamezne stroke, vendar skuša čim bolj celovito zajeti kompleksnost terminoloških besednih zvez, za namene tega eksperimenta pa je bilo luščenje omejeno le na dva oblikoskladenjska vzorca, in sicer enega enobesednega (*samostalnik*) ter enega dvobesednega (*pridevnik* + *samostalnik*). Nanju smo se omejili zato, ker za enobesedne *samostalniške* termine ter termine v obliki zveze *pridevnik* + *samostalnik* velja, da so v slovenskem jeziku strokovnopoimenovalno najbolj produktivni (Logar Berginc et al. 2013). Oblikoskladenjske vzorce smo poleg tega omejili zgolj na občne samostalniške zveze brez lastnoimenskih sestavin, iz nadaljnje obravnave pa smo izločili tudi vse izraze, ki so se v gradivnih korpusih pojavili manj kot petkrat.

2.2 Korpus

Terminološke kandidate smo luščili iz dveh korpusov, in sicer iz:

a) specializiranega korpusa glasbenih besedil (dalje glasbeni korpus) in

b) podkorpusa ccGigafide, v katerega smo združili besedila s temami iz biologije, kemije in fizike (dalje naravoslovni podkorpus).

Glasbeni korpus obsega deset učbenikov, ki so nastali v obdobju desetih let (2004–2014) ter se uporabljajo pri pouku glasbe na osnovni in srednji stopnji glasbenih šol. Korpus je nastal v okviru doktorske raziskave Jelene Grazio na Oddelku za

muzikologijo Filozofske fakultete Univerze v Ljubljani, učbeniki pa obravnavajo različne glasbene prvine: harmonijo (učbenika Harmonija I in Harmonija II Janeza Osredkarja), kontrapunkt (Kontrapunkt Janeza Osredkarja), oblikoslovje (Oblikoslovje Larise Vrhunc), solfeggio (Solfeggio I, II, III in IV Tomaža Habeta) ter teorijo glasbe v splošnih potezah (Sodobna teorija glasbe P. Amaliettija in Osnove glasbene teorije Pavla Mihelčiča).

Naravoslovni podkorpus je sestavljen izključno iz besedil, ki so zajeta v ccGigafidi (Logar Berginc et al. 2012: 77–97; Erjavec in Logar Berginc 2012). V njem je 13 osnovno- in srednješolskih učbenikov za naravoslovje, biologijo, fiziko ali kemijo ter še 16 drugih strokovnih in poljudnostrokovnih knjig različnih založb na temo astronomije, botanike in vrtnarjenja. Ostala besedila so iz naslednjih revij: Gaia, Gea, Kmetovalec, Moj lepi vrt, Moj mali svet, Mrgolazen, National Geographic, Revija o konjih, Ribič ter Rože in vrt.[3]

Osnovne podatke o obeh podkorpusih povzema Tabela 1.

Tabela 1: Osnovni podatki o glasbenem korpusu in naravoslovnem podkorpusu.

	Glasbeni korpus	Naravoslovni podkorpus
število pojavnic	280.060	1.053.897
število različnic	12.121	59.788
število dokumentov	10	388
vrsta besedil	učbeniki	učbeniki, poljudnostrokovne knjige in revije

3 ANALIZA REZULTATOV

Število izluščenih terminoloških kandidatov za posamezni korpus in oblikoskladenjski vzorec kaže Tabela 2, v Tabelah 3 in 4 pa so podani vrhnji deli vseh štirih seznamov.

Tabela 2: Število izluščenih kandidatov za posamezni korpus in oblikoskladenjski vzorec.

	Glasbeni korpus	Naravoslovni podkorpus
samostalnik	1.137	7.853
pridevnik + samostalnik	828	1.309

3 Področja biologija, kemija in fizika v Gigafidi z besedili niso enakovredno zastopana, kar je razvidno tudi iz izluščenih terminoloških kandidatov.

Iz Tabele 2 je razvidna razlika med obema podkorpusoma v velikosti, ki pri samostalniških kandidatih izkazuje podobno razmerje med potencialno terminološkimi samostalniki in vsemi različnicami (9,3 % pri glasbi in 13 % pri naravoslovju), pri dvobesednih izrazih pa prednjači glasba s 6,8 % proti naravoslovju z 2,2 %.

Tabela 3: Vrhnji del obeh seznamov izluščenih terminoloških kandidatov iz glasbenega korpusa.

Samostalnik	Pogostost v Gigafidi	Pridevnik + samostalnik	Pogostost v Gigafidi
1. ton	32.538	osnovni ton	201
2. akord	3.049	pesemska oblika	17
3. oblika	357.909	vodilni ton	13
4. glasba	290.529	sonatna oblika	14
5. interval	6.620	tonovski način	24
6. stavek	43.996	taktovski način	10
7. tema	231.129	alterirani akord	0
8. tonaliteta	305	akordično območje	0
9. takt	6.515	osnovna tonaliteta	2
10. kvinta	189	notna vrednost	15
11. glas	272.023	durova lestvica	13
12. nota	29.362	alikvotni ton	5
13. terca	439	zgornji glas	9
14. dur	4.037	cel ton	14
15. oktava	603	stranska stopnja	0
16. melodija	27.394	tripolovinski takt	0
17. stopnja	199.417	glasbena teorija	269
18. lestvica	138.848	alterirani ton	2
19. septima	35	menjalni ton	1
20. način	551.177	tonski način	56
21. trozvok	72	osnovna oblika	1.036
22. skladba	81.732	notno črtovje	238
23. primer	906.970	molova lestvica	17
24. gibanje	136531	dominantni septakord	12
25. d	61.852	uvajalna vaja	7
26. četverozvok	10	dominantni četverozvok	0
27. c	42.002	velika terca	16
28. vaja	92.132	lestvična stopnja	0
29. polovinka	101	akordični ton	1
30. kadenca	356	tonalni plan	1

Tabela 4: Vrhnji del obeh seznamov izluščenih terminoloških kandidatov iz naravoslovnega podkorpusa.

Samostalnik	Pogostost v Gigafidi	Pridevnik + samostalnik	Pogostost v Gigafidi
1. rastlina	132.625	ribiška družina	5.599
2. voda	516.116	ribolovna dovolilnica	962
3. vrt	144.814	sladkovodno ribištvo	1.134
4. RDA	11.876	živa meja	5.147
5. riba	101.596	ribiška zveza	1.185
6. vrsta	612.575	organsko gnojilo	1.913
7. konj	96.764	botanični vrt	3.886
8. ribič	34.728	velika količina	25.964
9. leto	4695.764	okrasna rastlina	2.899
10. tla	162.784	mlad ribič	1.257
11. cvet	47.693	ribiški čuvaj	802
12. list	197.567	veliko število	52.554
13. sorta	45.117	zelenjavni vrt	2.337
14. žival	202.023	različna vrsta	14.756
15. čas	1950.895	ribiški okoliš	993
16. cm	104.836	soška postrv	1.595
17. slika	570.703	sadno drevje	4.192
18. delo	1703.484	ribji živelj	713
19. ribolov	19.679	hlevski gnoj	4.530
20. zemlja	161.980	potočna postrv	1.212
21. kg	111.327	cerkniško jezero	3.431
22. barva	245.458	cvetlični lonček	1.412
23. drevo	84.742	športni ribolov	1.893
24. površina	152.514	organska snov	2.915
25. oblika	368.503	nizka temperatura	9.432
26. seme	42.020	nova sorta	1.615
27. jezero	88.267	sladkorna pesa	3.549
28. temperatura	110.203	ekološko kmetovanje	3.097
29. prostor	691.860	visoka temperatura	10.700
30. fotografija	237.996	članska izkaznica	1.337

Pogled na Tabeli 3 in 4 pokaže dve temeljni razliki med podkorpusoma, in sicer v ravni specializiranosti in homogenosti. Predvsem pri seznamu izluščenih izrazov iz naravoslovnega korpusa opazimo vpliv različnih (pre)močno zastopanih virov, denimo s področij ribištva in vrtnarstva. Poleg tega so očitne tudi velike razlike v

pogostosti v Gigafidi, še posebej pri dvobesednih izrazih; bolj specializirani glasbeni korpus vsebuje izraze, ki se v Gigafidi redkeje pojavljajo kot naravoslovni izrazi, ki so bili izluščeni iz podkorpusa Gigafide.

V postopku analize smo natančno pregledali le vrhnjih 150 enot na seznamih terminoloških kandidatov iz obeh korpusov. Analiza je potrdila pričakovano razliko v številu enot, ki bi jih iz prvega oz. drugega korpusa dali v splošno košarico. V obeh seznamih iz glasbenega korpusa je bilo primerov za splošno košarico veliko manj kot v seznamih iz naravoslovnega podkorpusa, natančneje: samostalniških kandidatov, ki bi jih zelo verjetno umestili v splošno košarico,[4] je v glasbenem korpusu v vrhnjem delu seznama le približno tretjina (*takt*, *glas*, *nota*, *melodija*, *skladba*, *harmonija* ipd.), pri zvezah pridevnika in samostalnika pa celo manj kot 15 % (npr. *notno črtovje*, *klasična glasba*, *klavirska spremljava*), medtem ko bi ostale izluščene enote iz glasbenega korpusa sodile v šolsko ali strokovno (npr. v strokovno: *modulacija*, *fuga*, *kvintakord*, *tritonusna kvinta*, *eolska septima*, *napolitanski sekstakord*).[5]

Na drugi strani je stanje pri seznamih iz naravoslovnega podkorpusa obratno: samostalnikov in zvez iz Gigafidinega podkorpusa, ki bi jih dali v splošno košarico, je velika večina, prim.: *rastlina*, *voda*, *list*, *seme*, *plod*, *poganjek*, *temperatura*, *svetloba*; *okrasna rastlina*, *soška postrv*, *organski odpadek*, *listna uš* itd. To pomeni, da v vrhnjem delu seznama izluščenih enot iz naravoslovnega podkorpusa skoraj ni leksike, ki bi jo bilo treba obravnavati ožje terminološko, sploh pa ne v smislu strokovne košarice. Manjši del enot iz Gigafide bi tako lahko šel le še v šolsko košarico – izmed prvih 300 enobesednih smo sem uvrstili besede: *celica*, *molekula*, *muha* ('ribištvo'), *beljakovina*, *dušik*, *masa*, *spojina*, *atom*, *kromosom*, *sila*, *populacija* in *pH*. Če s pregledom seznama nadaljujemo do 500. mesta, bi sem lahko prišteli še poimenovanja: *križanec*, *bala*, *bakterija*, *podtaknjenec*, *ozvezdje*, *aminokislina*, *membrana*, *humus*, *elektron*, *kasač*, *uplenitelj*, *herbicid*, *gen*, *insekticid* in *siliranje*. Groba ocena torej kaže, da je v seznamu izluščenih samostalniških terminoloških kandidatov okrog 5-odstotni delež poimenovanj, ki imajo potencial za obravnavo v šolski košarici. Med večbesednimi kandidati je takih več, tj. okrog 15 % med vrhnjimi 300 enotami, npr. *ogljikov hidrat*, *celična membrana*, *magnetno polje*, *maščobna kislina*, *potencialna energija*, *vrtilni moment*.

Analiza je torej dala naslednje odgovore na naši prvi dve raziskovalni vprašanji:

1. Metoda luščenja terminoloških kandidatov v ccGigafidi praktično ne izkaže gradiva, ki bi zahtevalo ozko terminološko obravnavo, izkaže pa tudi zelo malo gradiva, pri katerem bi morali leksikografi paziti na

4 Podajamo subjektivno oceno, ki zadošča za preliminarne ugotovitve, pred dokončnimi zaključki pa bi jo bilo treba potrditi še z večjim številom ocenjevalcev.

5 Ne bi jih seveda dejansko umeščali v slovar, v mislih imamo le primerjavo med seznami.

področno vezanost iztočnice (morda tudi oznako). Gigafida torej – vsaj pri naravoslovnih besedilih in v vrhnjem delu seznama – v veliki večini vsebuje le poimenovanja, pri katerih, kot smo navedli že zgoraj, sicer prepoznamo navezavo na določeno strokovno področje, vendar poznavanje področja ni pogoj za njeno razumevanje in pravilno rabo.

2. V nasprotju s preteklimi luščenji, ki smo jih izvedli na specializiranih korpusih strokovnih besedil, smo tokrat metodo preizkusili na področno heterogenem naboru poljudnostrokovnih besedil. Taka so namreč tipična besedila v splošnih korpusih: ubesedujejo teme, ki povezujejo različna področja, ter jih opisujejo in razlagajo na nestrokovnjakom prilagojen način. Eksperiment je pokazal, da je z vidika analize rezultatov prihodnja taka luščenja bolje načrtovati na tematsko kolikor se da enotnih besedilnih zbirkah, četudi bi se na ta način opredeljeni viri (npr. revija Gea) pojavljali v več izvedbah te metode. Za leksikografsko analizo je namreč manj moteče, če smo pri pregledu osredotočeni na poimenovanja (in njihovo ožjo terminološkost) zgolj enega področja. Obenem je treba poudariti, da bi za namene slovarja, ki naj bi zajemal tudi specializirano izrazje strok na ravni srednješolskih učbenikov, splošni korpus morali dopolniti z ustreznimi učbeniki, pri luščenju pa uporabiti primerjavo med homogenim specializiranim korpusom in splošnim korpusom brez stvarnih besedil.
3. Razvrščanje v košarice mora v končni implementaciji luščenja potekati samodejno. Čeprav se že v tu opisanem pilotnem eksperimentu zarisujejo frekvenčna območja, v katerih se gibljejo izrazi v splošni, šolski in strokovni košarici, je treba metodologijo razvrščanja šele razviti skupaj s spremenljivkami, ki bodo odvisne od posameznega strokovnega področja (npr. frekvenčni pragovi in oblikoskladenjski vzorci za luščenje). Namesto absolutne korpusne pogostosti bo metoda po vsej verjetnosti uporabljala pogostostno razmerje med strokovnim in splošnim delom korpusa oziroma navzkrižne primerjave med podkorpusi sorodnih področij. Temeljna ovira, da takšne metodologije v tem trenutku še ni mogoče predlagati, je neobstoj orodja za samodejno razdvoumljanje, ki bi omogočalo razdelitev korpusne pojavitve izrazov na posamezne pomene, šele s dostopnostjo podatka o pogostosti posameznega pomena pa lahko učinkovito (in tudi samodejno) presojamo o terminološkosti.

Ker je pri sedanji metodi luščenja za izračun terminološkosti uporabljena primerjava pogostosti med specializiranim in celotnim splošnim korpusom, se pri rezultatih jasno izkaže tudi nesorazmerna zastopanost področij v Gigafidi; tako se denimo pogostosti določenih leksikalnih enot lahko nesorazmerno povečajo zgolj zaradi določene revije, ki je v korpus vključena z več letniki. Čeprav se načrtuje

sistematično dopolnjevanje Gigafide z besedili tistih področij, ki so zdaj slabše zastopana, bi bilo iluzorno pričakovati, da bo uravnoteženost korpusa kdajkoli popolna; pravzaprav področna uravnoteženost niti ni cilj referenčnega korpusa.

4 SKLEP

Ob koncu velja še enkrat poudariti, da nas v eksperimentu ni zanimala ocena terminološke uspešnosti luščenj iz dveh korpusov – enega področno specializiranega in homogenega, drugega splošnejšega in heterogenega – v smislu terminografije, temveč smo si zadali nalogo s to metodo poiskati terminološkost v splošnem jeziku. Predvidevali smo, da bi preliminarni rezultati lahko izpostavili nekatere prednosti in slabosti ter nakazali nekatere smernice za prilagoditev pristopa, ki ima prvotno drugačen namen. Čeprav smo iz luščenja izpustili časopisje, internetna besedila in kategorijo »drugo«,[6] so seznami, pridobljeni po trenutni metodi luščenja iz heterogenega korpusa naravoslovja, izkazali majhen obseg enot, ki bi v prihodnjem slovarju slovenščine, nastalem na podlagi Gigafide, potrebovale terminološko pozornost, nasprotno sliko pa kaže glasbeni korpus. Analiza seznamov je pokazala tudi to, da bo za leksikografski postopek potrebna še nadaljnja natančnejša opredelitev košaric, pri nadgradnji korpusa pa temeljitejša dopolnitev z besedili, ki vsebujejo terminologijo, s katero se v času izobraževanja sreča velika večina govorcev slovenščine, tj. z učbeniki in sorodnim gradivom (Logar 2015).

6 Ta del Gigafide prav tako vsebuje determinologizirano leksiko, prim. npr. pravna poimenovanja *pogodba*, *odločba*, *sklep*, *pritožba*, *zahtevek*, značilna za spletni del korpusa, v Logar Berginc in Ljubešić (2013: 102).