Primerjava luščilnikov terminologije Sketch Engine in CollTerm za znanstvena besedila

Klara Eva Kukovičič

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
klara.eva@gmail.com

Povzetek

Namen članka je primerjati luščilnika terminologije SketchEngine in Collterm na primeru znanstvenih besedil. Natančna terminološka baza ali vsaj enojezični seznami terminov nam lahko pri procesu prevajanja znanstvenih in strokovnih besedil prihranijo veliko časa, zato je pomembno, da njihovo izdelavo vključimo v proces analize besedila oziroma predprevajanja. Samodejni luščilniki terminologije in kolokacij pri tem igrajo pomembno vlogo, saj je gradnja jezikovnih baz z njihovo pomočjo hitrejša in enostavnejša. V prispevku smo raziskovali, kateri izmed luščilnikov terminologije iz izbranih doktoratov iz Korpusa akademske slovenščine izlušči bolj relevantne termine. Pričakujemo, da bosta luščilnika izluščila različno število terminov, prav tako pa bo njihova razporeditev glede na ključnost različna.

Comparison of Sketch Engine and CollTerm extracion tools for scientific texts

The purpose of this article is to compare two tools for automatic term extraction, Sketch Engine and CollTerm for scientific texts. A termbase or monolingual lists of terms can save translators a lot of time in the process of translating scientific texts. For this reason, it is very important to include them in the process of text analysis or pre-translation. Here, automatic term extraction tools play an important role, as their functions help us creating bases faster and simpler. In this article, we will focus on two extraction tools and research which of the tools extract more for translators' relevant terminology from the Slovene corpus KAS.

1. Uvod

Z razvojem interneta se je prevajalski proces drastično spremenil. Tiskane slovarje in slovnice so nadomestili spletni viri, med katerimi so nepogrešljivi spletni slovarji, korpusi, tezavri, razni spletni portali, kot sta Fran[1] in Terminologišče[2] in mnogi drugi. V devetdesetih letih prejšnjega stoletja so se začela razvijati tudi prevajalska namizja, med katerimi velja omeniti vodilne SDL Trados Studio[3], MemoQ[4] in Memsource[5], MateCat[6]. Njihova skupna značilnost je pomnilnik prevodov, običajno pa ta orodja ponujajo še podporo za prevajanje različnih datotečnih formatov, vtičnik za strojno prevajanje, orodje za samodejno preverjanje kakovosti, orodja za vzporejanje dokumentov in seveda upravljanje terminologije.

Prevajalci porabijo približno 30-60 % časa prevajanja zgolj za iskanje ustrezne terminologije (Gornostay et al, 2010). Ravno zato je za prevajalce zelo pomembna terminološka baza, v katero lahko vnaprej ali sproti vpisujejo terminološke vnose. Pomemben korak pri izdelavi terminološke baze predstavlja priprava enojezične baze terminov, ki skrajša čas izdelave terminološke baze (Ponikvar, 2002: 19). Enojezično bazo terminov pa najenostavneje naredimo z orodjem za luščenje terminologije, kot so na primer SketchEngine, CollTerm, LUIZ, Lexterm, SDL MultiTerm Extract.

Različni luščilniki terminologije lahko temeljijo na jezikoslovnem, statističnem ali hibridnem pristopu, ki je kombinacija prvih dveh. Prav tako ima vsak luščilnik vgrajena pravila, po katerih izlušči termine.

V tem prispevku bomo z uporabniškega vidika primerjali dva luščilnika terminologije, in sicer SketchEngine in Collterm, na kratko predstavili raziskovalno področje, potek in rezultate analize ter težave, na katere smo naleteli ob analizi.

2. Namen članka

Namen tega prispevka je na kratko predstaviti Korpus akademske slovenščine (KAS), iz katerega smo črpali gradivo za analizo, področje terminologije, luščenja terminov in orodij za samodejno luščenje terminologije, Sketch Engine in Collterm, ter ugotoviti, katero izmed orodij ponudi boljši oziroma uporabnejši nabor terminov. Naročniki prevodov običajno težijo k čim prejšnji oddaji prevoda, zato smo prevajalci pogosto pod časovnim pritiskom in si ne vzamemo dovolj časa, da bi pregledali nabor vseh terminoloških izpisov, ki nam jih ponudijo luščilniki. Posledično smo se tudi v naši raziskavi odločili, da se osredotočimo zgolj na prvih 100 kandidatov, ki sta jih iz izbranih doktoratov Korpusa akademske slovenščine izluščila Sketch Engine in CollTerm.

3. Korpus akademske slovenščine

Projekt »Slovenska znanstvena besedila: viri in opisi« se je začel leta 2016 kot odgovor na cilje Akcijskega načrta za jezikovno izobraževanje (2015) in Akcijskega načrta za jezikovno opremljenost (2015), ki sta ugotovila, da je potrebno razvijati slovenščino v visokem šolstvu in znanost ter izboljšati položaj slovenščine kot jezika znanosti. V okviru projekta je bil ustvarjen korpus pisnih besedil akademske slovenščine (Erjavec et al.,: 2016).

Korpus akademske slovenščine vsebuje skoraj 1,2 milijarde pojavnic, največji delež korpusa (81,14 %) predstavljajo diplomska dela, sledijo magistrska dela (12,60 %) in doktorska dela (1,38 %) (Erjavec et al., 2016).

[1] https://fran.si/
[2] https://isjfr.zrc-sazu.si/terminologisce#v
[3] https://www.sdltrados.com/
[4] https://www.memoq.com/en/
[5] https://www.memsource.com/
[6] https://www.matecat.com/

V korpusu so najbolj zastopana besedila iz družboslovja (57 %), tehnoloških ved (32 %) ter naravoslovno-matematičnih ved (9 %) (Erjavec et al., 2016).

4. Termini in terminologija

Ključni element znanstvenih in akademskih besedil predstavlja terminologija, s katero se v prvi vrsti ukvarjajo terminologi in lingvisti (Fišer et al., 2016: 35). Po načelih terminološke vede naj bi bili termini enote, ki so pomensko predvidljive in ustaljene. Z razvojem računalniškega pristopa k pridobivanju terminologije pa se je izkazalo, da se termini dinamično spreminjajo v odvisnosti od besedilnih dejavnikov in da niso skladenjsko in oblikoslovno usklajene enote (Vintar, 2009: 347). Termine lahko definiramo tudi kot jezikovne znake, ki označujejo pojem in se po svojih lastnostih razlikujejo od druge leksike (Fajfar et al., 2015).

4.1. Subjektivno dojemanje terminološkosti

O subjektivnosti dojemanja terminologije je bilo do sedaj narejenih že več raziskav. Ena izmed obsežnejših je bila izvedena v okviru doktorske dizertacije, kjer je R. Estopà Bagot analizira, kako različne skupine uporabnikov terminologije dojemajo terminološkost (Vintar, 2008: 47-49). Rezultati so pokazali velika odstopanja med prevajalci, dokumentalisti, strokovnjaki in terminografi (Vintar, 2008: 47-49). Največ enot so za termine označili terminografi (1052), strokovnjaki so označili 938 izrazov, dokumentalisti 486 izrazov, najmanj enot pa so kot termine označili prevajalci (270) (Vintar, 2008: 47-49).

Podobne raziskave pa so bile narejene tudi v slovenskem prostoru (Fajfar et al., 2015: 8). Rezultati raziskave, ki je potekala v okviru projekta TERMIS, so pokazali, da so študenti v analiziranih besedilih označili manj besed za termine, kot so pričakovali avtorji raziskave (Logar Berginc 2013: 248). Avtorice druge raziskave, v katero sta bila vključena dva strokovnjaka s področja odnosov z javnostmi, pa v sklepu prav tako ugotavljajo, da neujemanje rezultatov obeh udeležencev »jasno kaže subjektivnost same definicije terminološkosti« (Logar Berginc et al., 2013: 132-133).

4.2. Orodja za samodejnost luščenje terminologije

Vintar (2009: 345) samodejno luščenje terminologije definira kot postopek, pri katerem program na podlagi statističnih izračunov, jezikoslovnih analiz ali drugih obstoječih podatkov skuša ugotoviti, katere besede in besedne zveze v danem korpusu strokovnih besedil so terminološke. Področje procesiranja naravnih jezikov se že več kot 20 let aktivno ukvarja z luščenjem terminologije, ki lahko temelji na statistični ali jezikovni metodi ali kombinaciji obeh (Pinnis et al., 2012: 193), danes pa se s pridom uporabljajo tudi metode strojnega učenja.

4.2.1. Sketch Engine

Sketch Engine je spletno orodje, ki uporabnikom omogoča brskanje in analiziranje že obstoječih in ustvarjanje lastnih korpusov (Fišer et al., 2016: 136). Orodje prav tako omogoča samodejno luščenje terminologije. Za luščenje terminologije z orodjem Sketch Engine potrebujemo dva korpusa, in sicer prvega, iz katerega želimo izluščiti kandidate, in drugega, tj. referenčnega korpusa, ki mora biti čim večji. Orodje nato primerja besedišče prvega korpusa z besediščem drugega, referenčnega korpusa, in na podlagi pravil izlušči kandidate.

Proces luščenja z orodjem Sketch Engine je dvostopenjski. Prvi korak temelji na pravilih in je odvisen od jezika. V tem koraku se z uporabo tako imenovane slovnice terminov (ang. term grammar) oceni slovnična veljavnost določene zveze v specializiranem korpusu. V drugem koraku pa se kandidate, ki smo jih dobili v prvem koraku, primerja z referenčnim korpusom z uporabo »simplemath« statistike (Fišer et al, 2016: 36). Slovnico terminov za slovenščino so na podlagi češke predloge razvili Fišer et al. (2016). Za slovnico terminov uporabljamo poizvedovalni jezik Corpus Query Language (CQL). Zapis slovnice terminov si lahko ogledamo na naslednjih primerih. Najprej je zapisan primer slovnice za dvobesedne termine, nato pa še za tri- in štiribesedne:

```
1:adj 2:noun & agree(1,2)
*COLLOC "ll_%(1.gender_lemma)_%(2.lemma_lc)-x"
```

V prvi vrstici je pravilo, ki ga orodje uporablja za prepoznavanje vzorca, v drugi pa pravilo, kako naj se prepoznani termin izpiše. Primer ponazarja prepoznavo dvobesednih zvez pridevnika in samostalnika, v katerih se spol pridevnika ujema s spolom samostalnika. Program prepozna vse besedne zveze v korpusu, ki ustrezajo temu pravilu. Nato s pomočjo statistike izvede rangiranje prepoznanih zvez glede na rezultate njihove terminološkosti in jih izpiše v skladu s pravilom v drugi vrstici. Za zgornji primer je navodilo tako, da se samostalnik izpiše kot lema z malo začetnico, pridevnik pa v lemi, ki se po spolu ujema s samostalnikom. Primer izluščenega termina, ki ustreza temu pravilu, je npr. *prosti delec.*

Primer slovnice terminov za tribesedne termine:

```
1:adj 2:noun 3:noun_genitive
*COLLOC
"z_%(1.gender_lemma)_%(2.lemma_lc)_%(3.lc)-x"
```

Zgoraj zapisani primer torej pove, da program v seznamu kandidatov izpiše termine, ki so sestavljeni iz zveze pridevnika, samostalnika in samostalnika v rodilniku, pri čemer se spol pridevnika ujema s spolom samostalnika. Primer takšnega termina je na primer lahko *samodejno prepoznavanje glasu*.

Primer slovnice terminov za štiribesedne termine:

```
1:noun 2:adj_genitive 3:noun_genitive 4:noun_genitive &
agree(2,3)
*COLLOC "d_%(1.lemma_lc)_%(2.lc)_%(3.lc)_%(4.lc)-
x"
```

Primer štiribesednega termina, ki ustreza zgornjemu pravilu, je npr. *gostota prostih nosilcev naboja*.

4.2.2. CollTerm

O CollTerm je orodje za luščenje kolokacij in terminologije, ki je bilo razvito v okviru projekta ACCURAT in je prosto dostopno. Orodje so razvili Pinnis et al. (2012). Predhodna različica orodja je terminološke kandidate luščila zgolj s pomočjo oblikoskladenjskih

vzorcev in z različnimi statistikami za sopojavitev besed oziroma besednih zvez. Danes je orodje nadgrajeno z modulom za nadzorovano učenje. To omogoča, da orodje na koncu poda en sam seznam rangiranih terminoloških kandidatov in ne po en seznam za vsako stopnjo n-gramov kot prej.

CollTerm uporablja enake oblikoskladenjske vzorce kot Sketch Engine, prilagojene za slovenščino. Proces luščenja poteka v štirih korakih, in sicer:

1. morfosintaktično filtriranje,
2. filtriranje za minimalno pogostost
3. statistično rangiranje glede na referenčni korpus in
4. filtriranje na podlagi rezultata za rangiranje oz. uvrstitve na rangiranem seznamu.

Orodje CollTerm tako kot Sketch Enigne omogoča, da shranimo izpis terminoloških kandidatov, kjer sta razvidni tako ključnost kot tudi pogostost pojavitve. Višja kot je ključnost, bolj relevanten je za nas izpisan terminološki kandidat.

Orodje CollTerm za razliko od Sketch Engina pri izpisu terminoloških kandidatov zapiše obliko kandidata v prvem sklonu in v sklonu, v katerem se največkrat pojavi. Če se torej kandidat največkrat pojavlja v četrtem sklonu (npr. umetno inteligenco), ga orodje izpiše tako v tožilniku kot tudi v imenovalniku, pri čemer pa lahko pride do neujemanja spola samostalnika s pridevnikom (npr. umeten inteligenca).

Obe orodji rangirata kandidate glede na »keyness score« oziroma ključnost. Višja kot je ključnost, višje se pojavi kandidat.

5. Zasnova projekta

Delo smo začeli s pregledom doktoratov, vključenih v korpus KAS. Želeli smo izbrati terminološko in tematsko čim bolj podobne doktorate, zato smo se odločili, da se osredotočimo na področje računalništva. S seznama več kot sedemsto doktoratov smo nato izbrali vse tiste, ki so bili napisani na Fakulteti za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru, saj korpus ne vsebuje doktoratov s Fakultete za računalništvo in informatiko, Univerze v Ljubljani. Ker je bil tudi teh doktoratov preveliko, smo jih nato izbrali glede na tematiko, o kateri so pisali. Za temo smo določili umetno inteligenco in dobili precej manjši nabor, iz katerega smo nato določili tri doktorate, ki so si bili vsebinsko najbolj podobni.

Za vsak izbrani doktorat iz KAS je nato sledila priprava podkorpusa, ki smo ga izdelali z orodjem Sketch Engine. Sledil je proces luščenja terminoloških kandidatov, ki je potekal anonimno, saj smo želeli pri analizi ohraniti čim večjo objektivnost. Pri luščenju terminologije smo se obrnili na doc. dr. Darjo Fišer, Filozofska fakulteta Univerze v Ljubljani, in dr. Nikolo Ljubešića, Institut Jožef Štefan, ki sta z orodjem Sketch Engine in CollTerm izluščila kandidate treh izbranih doktoratov. Pri orodju Sketch Engine smo pri možnostih za izpis terminoloških kandidatov iz podkorpusov izbrali možnost »Keywords« oziroma ključne besede, kot referenčni korpus pa smo tako pri obeh orodjih določili celotni korpus KAS. Ker se je izkazalo, da Sketch Engine omogoča luščenje zgolj večbesednih kandidatov, smo tudi v orodju CollTerm omejili izpis na večbesednih kandidatov.

Seznami s terminološkimi kandidati niso vsebovali podatkov o orodju, ki je bilo uporabljeno, s čimer smo pri evalvaciji ohranili objektivnost. Sezname kandidatov smo na začetku označili kot Seznam orodja 1 in Seznam orodja 2. Šele po koncu analize smo preverili, katero orodje je bilo zapisano pod številko 1 in katero pod številko 2.

Naslednji korak je bilo označevanje kandidatov. Pri tem smo uporabljali Navodila za ocenjevanje terminoloških kandidatov, ki so v projektu KAS že služila kot pripomoček in navodilo za označevanje terminov. Smernice vsebujejo pet kategorij, in sicer termin, izvenpodročni termin, nerelevantno, znanstveno pisanje in »ne vem«. Kategorija termin vključuje besede in besedne zveze, ki predstavljajo termine z določenega področja. V kategorijo izvenpodročni termini so bile uvrščene besede in besedne zveze, ki so po oceni označevalca termini z drugega področja. Kategorija znanstveno pisanje zajema izrazje, vezano na pridobivanje, analizo ali predstavitev podatkov v doktorski raziskavi (npr. *tabela*) in besedišče, ki je stalni del doktorskega pisanja ali strukture doktorskega dela (npr. *zaključek*). V kategorijo nerelevantno uvrščamo splošno besedišče (npr. *celota, v bistvu*), zveze, ki so daljši opisi ali ki zahtevajo razpravljalno definicijo oz. so subjektivne narave in besede ali besedne zveze, ki so le del termina ali pa poleg termina vsebujejo še druge elemente (npr. *model govorca* – prekratko, termin je namreč *univerzalni model govorca* in *grozdenje govorcev* – predolgo, termin je namreč zgolj *grozdenje*). Označevalec kandidate po potrebi preverja v kontekstu korpusa KAS s pomočjo konkordanc. Kandidate, ki jih tudi s pomočjo konteksta ni mogoče uvrstiti v eno od kategorij, označimo z oznako »ne vem«.

Z uporabo istih smernic tudi v pričujoči raziskavi smo zagotovili večjo konsistentnost vrednotenja terminov. Izpustili smo zgolj kategorijo »ne vem«, saj bi ta kategorija izkrivila rezultate naše analize. Evalvacijo luščilnikov terminologije sem opravila avtorica prispevka, podiplomska študentka prevajanja.

Kot pomoč pri odločanju, ali je izluščena beseda oz. besedna zveza termin s področja računalništva ali informatike, pa so mi poleg lastnega znanja služili naslednji viri:

a) Islovar: terminološki, razlagalni in informativni slovar, ki strokovno izrazje pomensko in jezikovno opisuje in vrednoti. Zajema izrazje informatike, informacijske tehnologije in telekomunikacij ter drugih računalniških področij. Slovar ne vsebuje besed splošnega izrazja (Islovar).
b) DIS slovarček: slovar računalniških izrazov, verzija 2.1.71, ki je nastal na oddelku za inteligentne sisteme na Inštitutu Jožef Štefan. Slovarček trenutno vsebuje 12.296 gesel (DIS slovarček).
c) English/Slovene dictionary of computer science: spletni slovar računalništva, ki je nastal v okviru Inštituta Jožef Štefan, na voljo je tudi v tiskani verziji (English/Slovene dictionary of computer science).
d) Wikipedia: prosta spletna enciklopedija, ki vsebuje več kot 160.751 člankov (Wikipedia).
e) Prostodostopni zapiski in literatura s predavanj s Fakultete za računalništvo in informatiko.

Če smo na katerem koli viru našli definicijo terminološkega kandidata ali v slovenščini ali v angleščini, smo sklepali, da

gre za termin. Glavni namen označevanja kandidata je bil priprava podatkov za analizo, v kateri smo preverjali, kateri luščilnik terminologije je izluščil več kakovostnih terminov oziroma jih je rangiral višje. Že pred začetkom analize smo pričakovali, da bosta luščilnika podala različne kandidate, zato smo se želeli osredotočiti tudi na to, ali katero izmed orodij izlušči kakšen termin, ki ga drugi ne. Končni cilj je bil predvsem preveriti, katero izmed orodij tako profesionalnim prevajalcem kot študentom na prvih stotih mestih ponudi bolj relevantne terminološke kandidate.

6. Analiza

Analiza je sestavljena iz treh delov. V prvem delu nas je zanimala zgolj frekvenca posameznih oznak za posamezen seznam izluščenih kandidatov, v drugem delu smo računali odstotek pozitivnih primerov med prvimi N kandidati, v tretjem pa smo na primeru enega izmed doktoratov preverjali, ali so termini, izluščeni s prvim luščilnikom in ki se nahajajo med prvimi stotimi kandidati prvega seznama, vključeni tudi na seznam kandidatov, izluščenih z drugim luščilnikom.

6.1. Rezultati izluščenih kandidatov

Najprej smo prešteli, koliko izluščenih terminoloških kandidatov smo označili z oznako termin, izvenpodročni termin, znanstveno izrazje ali nerelevantno.

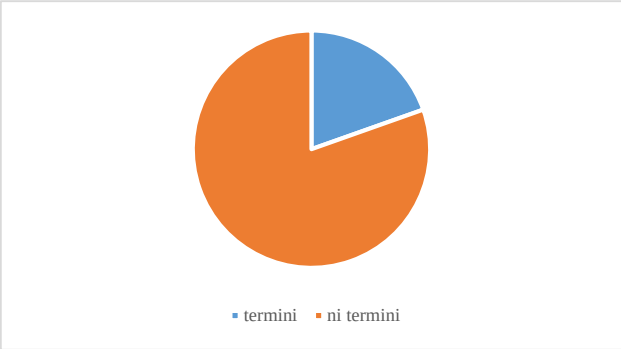
Kot je razvidno iz podatkov analize, je luščilnik CollTerm izluščil več terminov. Pri tretjem doktoratu, kjer smo s Sketch Enginom izluščili več terminov kot s CollTermom, je CollTerm izluščil bistveno več izvenpodročnih terminov. Pri drugih dveh doktoratih sta oba luščilnika izluščila enako število izvenpodročnih terminov, vendar ti niso bili isti.

	Oznaka	CollTerm	Sketch Engine
1	Termin	68	54
2	Izvenpodročni termin	13	3
3	Znanstveno pisanje	106	130
4	Nerelevantno	113	113

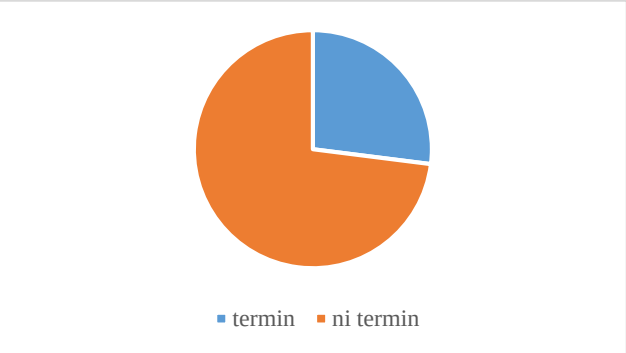
Tabela 1: Frekvenca oznak pri terminoloških kandidatih, izluščenih iz doktoratov

		CollTerm	Sketch Engine
1	Termini	81 (27 %)	57 (19 %)
2	Ne termini	219 (73 %)	243 (81 %)

Tabela 2: Število izluščenih terminov



Graf 1: Luščenje z orodjem Sketch Engine



Graf 2: Luščenje z orodjem CollTerm

Druga tabela prikazuje število in odstotek izluščenih terminov in ostalega besedišča. Vidimo, da je bilo 27 % terminoloških kandidatov, izluščenih s CollTermom, terminov. Nekoliko slabši rezultat smo dobili z orodjem Sketch Engine. Rezultati so predstavljeni tudi na Grafu 1 in 2. Graf 1 prikazuje razmerje med termini in ne termini, ki jih je izluščil Sketch Engine, Graf 2 pa razmerje, ki ga je izluščil CollTerm.

6.2. Odstotek pozitivnih primerov med prvimi N kandidati

V drugem delu analize smo se osredotočili na računanje odstotkov pozitivnih primerov med prvimi N kandidati. Kot pozitivne primere smo označili vse kandidate, ki smo jih pripisali oznako t (termin) ali x (izvenpodročni termin).

Namen drugega dela analize je bil, da ugotovimo, ali se termini na seznamu izluščenih kandidatov gosteje nahajajo na prvih mestih ali so enakomerno razporejeni čez celoten seznam. Ta del analize se nam je zdel ključen predvsem zato, ker se zavedamo, da pri prevajanju pogosto pripravimo zgolj ožji nabor terminologije, pri čemer v terminološko bazo vključimo prvih nekaj kandidatov.

6.2.1. Odstotek pozitivnih primerov med prvimi 20 kandidati

Spodnje tabele (Tabela 3, Tabela 4 in Tabela 5) prikazujejo odstotek terminov med prvimi dvajsetimi izluščenimi kandidati pri vsakem seznamu posebej. Če pogledamo povprečje vseh rezultatov, ki je prikazano v Tabeli 6, vidimo, da je med prvimi 20 kandidati, izluščenimi z orodjem CollTerm, 38 % kandidatov terminov, odstotek pa ni bistveno nižji (35 %) pri kandidatih, izluščenih s Sketch Enginom.

	Seznam 1
CollTerm	25 %
Sketch Engine	15 %

Tabela 3: Pozitivni primeri prvega seznama med prvimi 20 kandidati

	Seznam 2
CollTerm	45 %
Sketch Engine	50 %

Tabela 4: Pozitivni primeri drugega seznama med prvimi 20 kandidati

	Seznam 3
CollTerm	45 %
Sketch Engine	40 %

Tabela 5: Pozitivni primeri prvega seznama med prvimi 20 kandidati

	Povprečje vseh seznamov
CollTerm	38 %
Sketch Engine	35 %

Tabela 6: Povprečje pozitivnih primerov med prvimi 20 kandidati iz vseh seznamov

6.2.2. Odstotek pozitivnih primerov med prvimi 50 kandidati

Spodnje tabele (Tabela 7, Tabela 8 in Tabela 9) prikazujejo odstotek terminov med prvimi petdesetimi izluščenimi kandidati pri vsakem seznamu posebej. Povprečje vseh treh seznamov je prikazano v Tabeli 9. Rezultati kažejo, da se je razlika med odstotkom terminov, izluščenih z orodjema, v primerjavi z naborom 20 kandidatov, povečala. Odstopanje pri naboru prvih 20 kandidatov je bilo zgolj 3-odstotno, razlika pri naboru prvih 50 kandidatov pa se je povečala na 9 %.

	Seznam 1
CollTerm	28 %
Sketch Engine	12 %

Tabela 7: Pozitivni primeri prvega seznama med prvimi 50 kandidati

	Seznam 2
CollTerm	30 %
Sketch Engine	30 %

Tabela 8: Pozitivni primeri drugega seznama med prvimi 50 kandidati

	Seznam 3
CollTerm	36 %
Sketch Engine	24 %

Tabela 9: Pozitivni primeri tretjega seznama med prvimi 50 kandidati

	Povprečje vseh seznamov
CollTerm	31 %
Sketch Engine	22 %

Tabela 10: Povprečje pozitivnih primerov med prvimi 50 kandidati z vseh seznamov

6.3. Rangiranje izluščenih terminov

V tretjem delu analize smo se osredotočili zgolj na en doktorat. Najprej smo izpisali termine in njihovo zaporedno številko na seznamu kandidatov, ki je bil narejen z orodjem CollTerm. Nato smo preverili, pod katero zaporedno številko je isti termin rangiran na seznamu terminoloških kandidatov, ustvarjenim z orodjem Sketch Engine. Nato smo postopek zamenjali in izpisali termine in zaporedne številke terminov iz seznama kandidatov, narejenega z orodjem Sketch Engine in zaporedno številko primerjali z rangiranjem termina na seznamu kandidatov luščilnika CollTerm.

V Tabeli 11 lahko v levi koloni vidimo številko, pod katero je bil na seznamu CollTerm izluščenih kandidatov zapisan termin. Če je na primer zapisana številka 2, to pomeni, da je drugi zaporedni kandidat s seznama, ustvarjenega z orodjem CollTerm, termin. Desni stolpec za prikazuje, na katerem mestu se je isti termin pojavil na seznamu kandidatov, izluščenih z orodjem Sketch Engine. Tabela 12 pa prikazuje ravno obratno situacijo.

Namen zadnjega postopka analize je bil, da preverimo, ali sta luščilnika izluščila iste termine in kje so ti rangirani.

Zaporedna št. v CollT	Zaporedna št. v Sk E
2	30
5	21
7	9
13	205
16	67
17	0
18	135
19	**10**
21	4541
22	0
24	54
31	544
32	546
36	4727
39	40
40	4440
44	4422
50	4681
59	**7**

63	201
70	4669
71	454
72	4106
83	4594
88	0

Tabela 11: Rangiranje terminov

Zaporedna št. v Sk E	Zaporedna št. CollT
2	**185**
7	**59**
9	7
10	**19**
11	**192**
20	**195**
21	5
29	**194**
33	2
34	0
40	39
54	24
66	0
67	16
70	**219**
96	**215**

Tabela 12: Rangiranje terminov

V 11. in 12. tabeli so krepko označeni tisti termini, ki se prej pojavijo na seznamu terminoloških kandidatov, izluščenih s Sketch Enginom kot s CollTermom. Vidimo, da je sicer tudi Sketch Engine izluščil iste termine kot CollTerm, vendar so ti rangirani občutno nižje. Prav tako pa sta obe orodji izluščili nekaj terminov (Sketch Engine 2, CollTerm 3), ki jih drugo orodje ni.

Zahvala

Raziskavo, opisano v prispevku, je podprl projekt ARRS J6-7094 »Slovenska znanstvena besedila: viri in opis«.

7. Zaključek

Rezultati analize so pokazali, da se na seznamu terminoloških kandidatov obeh orodij nahaja precejšnje število terminov. Čeprav sta obe orodji v povprečju izluščili približno tretjino terminov, je orodje CollTerm pokazalo boljše rezultate oziroma je izluščilo več izrazja, relevantnega za prevajalce. CollTerm je v povprečju izluščil več terminov, ne samo med prvimi dvajsetimi in petdesetimi, ampak tudi med prvimi sto kandidati. Ob tem pa seveda moramo upoštevati, da je zaradi subjektivnega dojemanja terminologije verjetno prišlo do manjših odstopanj pri označevanju kandidatov s kategorijami: termin, izvenpodročni term, znanstveno pisanje in nerelevantno, čemur bi se v prihodnjih raziskavah dalo v dobršni meri izogniti z uporabo več označevalcev.

Analiza je prav tako pokazala, da so nekateri termini, ki nam jih ponudi CollTerm, na seznamu kandidatov, ustvarjenim z luščilnikom Sketch Engine, rangirani precej nizko ali pa jih na seznamu sploh ni.

V analizo nismo vključili podatka o tem, koliko terminov sta obe orodji spregledali (recall). Če bi želeli dobiti te podatke, bi morali ročno pregledati in označiti vsaj vzorec besedil, kar sicer presega okvire te raziskave, je pa v naših načrtih za prihodnje delo.

Poznavanje obeh orodij vsekakor nudi odlično podporo pri prevajalskem procesu. Čeprav so rezultati naše raziskave pokazali, da CollTerm prevajalcem ponudi ustreznejše jezikovne elemente, je odločitev, katero orodje bo prevajalec izbral, odvisna tudi od drugih faktorjev, kot sta na primer dostopnost in enostavnost uporabe, kjer ima orodje Sketch Engine v tem trenutku nekaj pomembnih prednosti.

8. Literatura

Akcijski načrt za jezikovno izobraževanje. 2015. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/raziskave-analize/slovenski_jezik/Akcijska_nacrta/ANJI.pdf (Dostop: 25. 3. 2018).

Akcijski načrt za jezikovno opremljenost. 2015. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Razpisi/2017/JR-ESS-Ranljive_skupine_govorcev/Akcijski_nacrt_za_jezikovno_opremljenost.pdf (Dostop: 25. 3. 2018).

DIS slovarček http://dis-slovarcek.ijs.si/ (Dostop: 10. 4. 2018).

Darja Fišer, Vit Suchomel, Miloš Jakubiček. 2016. Terminology Extraction for Academic Slovene Using Sketch Engine. RASLAN 2016.

English/Slovene dictionary of computer science https://www.ijs.si/cgi-bin/rac-slovar? (Dostop: 10. 4. 2018).

Islovar. http://www.islovar.org/islovar (Dostop: 10. 4. 2018).

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatiana Gornostay. 2012. Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. Proceedings of RASLAN 2016.

Meselina Ponikvar. 2002. Računalniška podpora prevajalskemu in terminološkemu delu na primeru prevajanja v okolju sektorja za prevajanje SVEZ. Diplomsko delo, Ljubljana, Univerza v Ljubljani, Filozofska fakulteta.

Nataša Logar Berginc, Špela Vintar, Špela Arhar Holdt, Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka, Slovenščina 2.0 1 (2013), št. 2 = Jezikovne tehnologije, ur. Tomaž Erjavec – Jerneja Žganec Gros, 113-138.

Sketch Engine Documentation https://www.sketchengine.eu/documentation/writing-term-grammar/#file (Dostop: 10. 8. 2018).

Špela Vintar. 2009. Samodejno luščenje terminologije – izkušnje in perspektive. V Terminologija in sodobna terminografija, ur. Nina Ledinek, Mojca Žagar in Marjeta Humar, str. 345-356. Ljubljana, Založba ZRC, ZRC SAZU.

Tanja Fajfar, Mojca Žagar Karer: Strokovnjaki in prepoznavanje terminov v strokovnih besedilih. Jezikoslovni zapiski, 21/1, 2015, str. 7-21.

Tatiana Gornostay, Andrejs Vasiljevs, Signe Rirdance, Roberts Rozins. 2010. Bridging the Gap – EuroTermBank Terminology Delivered to User´s Environment, Proceedings of the 14th Annual Conference of the European Association for Machine Translation.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, Milan Ojsteršek. 2016. Slovenska akademska besedila: prototipni korpus in načrt analiz. Zbornik konference Jezikovne tehnologije in digitalna humanistika https://core.ac.uk/download/pdf/143471580.pdf (Dostop 5. 4. 2018).

Wikipedia https://sl.wikipedia.org/ (Dostop: 10. 4. 2018).