Luščenje borzne terminologije

Senja Pollak*, Biljana Božinovski‡,

* Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova 39, 1000 Ljubljana
senja.pollak@ijs.si
‡ Biljana Božinovski
Maistrova ulica 2, 8250 Brežice
bbozinovski@poslovniprevodi.si

Povzetek

V članku je predstavljen pristop h gradnji geslovnika za slovar borzne terminologije, izdelan na podlagi avtomatskega luščenja terminologije. Predstavimo korpus slovenskega borznega jezika ter motiviramo izbiro pristopa luščenja z orodjem LUIZ-CF, ki ga tudi primerjamo s pristopom, temelječim na orodju WordSmith Tools. Izračunamo natančnost in priklic avtomatskega luščenja ter strinjanje med ocenjevalcema. V nadaljevanju analiziramo izluščene termine ter podamo predloge za izboljšavo luščilnika.

Extraction of Stock Market Terminology

The paper presents an approach to building the wordlist for a dictionary of stock market terminology using automatic terminology extraction. A specialised corpus of Slovene stock market terminology is presented, followed by the argumentation why terminology was extracted with LUIZ-CF. In addition, the approach to building a wordlist using LUIZ-CF is compared with the approach using WordSmith Tools. Precision and recall are calculated for both term extraction methods, and the level of agreement between two evaluators is examined. We analyse the extracted terminology and propose improvements of the selected term extraction tool.

1. Uvod

Inventar terminologije strokovnega področja je osnova za izdelavo terminološkega slovarja oziroma terminološke zbirke, ki je ključni jezikovni vir strokovnega prevajalca. Izdelava geslovnika lahko poteka ročno, kar zahteva veliko strokovnega znanja, je časovno zamudno ter izhaja iz subjektivnih izbir. Zato so se v zadnjem desetletju raziskovalci s področja računalniškega jezikoslovja posvetili izdelavi metod za avtomatsko luščenje terminologije iz korpusov. Samodejne metode so bile razvite za različne jezike, npr. za angleščino Sclano in Velardi (2007), Ahmad idr. (2007), Frantzi in Ananiadou (1999), Kozakov idr. (2004); za slovenščino rešitve ponuja Vintar (2003, 2010). Dvo- in večjezične rešitve predstavljajo npr. Lefever idr. (2009), Macken idr. (2013), na voljo pa so tudi plačljiva orodja, kot so SDL MultiTerm Extract,[1] WordSmith Tools[2] in SketchEngine.[3]

Namen članka je opisati uporabo jezikovnih tehnologij, natančneje avtomatskega luščenja terminologije, pri izdelavi geslovnika slovenske borzne terminologije. Angleška borzna terminologija je dostopna v slovarjih (npr. Barron's Dictionary of Finance and Investment Terms) in zlasti na spletu, kjer svetovne borze (npr. ameriški NASDAQ, angleški LSE, kanadski TMX, avstralski ASX) in investicijski portali (Investopedia, Investor Words) predstavljajo tudi najnovejše strokovne izraze. Slovenska borzna terminologija je samostojno predstavljena v Borznih izrazih (Čas in Rotar, 1994); gre za večjezični slovar, ki je služil kot podlaga za vključitev borznih izrazov tudi v številne kasnejše spletne terminološke zbirke finančnih institucij in investicijskih portalov (NLB, Abanka, vzajemci.com in številni drugi). Omenjene zbirke, ki sicer pokrivajo širša področja financ, bančništva in podobno, imajo s terminološkega in terminografskega vidika nekaj pomanjkljivosti: vsebujejo borzne izraze, ki niso več v uporabi (npr. francoska tujka *fond*), ne vsebujejo številnih na novo skovanih izrazov, ki so se v slovenski borzni terminologiji ustalili zlasti od začetka finančne krize naprej (npr. *slaba banka*), ter niso izdelane v skladu z načeli terminografske stroke (zapisi terminov z velikimi začetnicami/tiskanimi črkami, ciklične/nepopolne/netočne definicije ipd.). Zaradi odsotnosti standarda in neenotne rabe se v besedilih pojavljajo številne dvojnice (*investitor/vlagatelj, borzna kotacija/uradna kotacija, tečaj/cena vrednostnega papirja*), ki nestrokovnjaka begajo in otežujejo prevajanje. Kaže se potreba po sistematični analizi sodobnih slovenskih besedil z borznega področja in zajemu aktualnega izrazja (ter drugih besedilnih informacij) v dejanski rabi, in sicer za nadgradnjo obstoječih oziroma izdelavo normativnega dvojezičnega slovarja, pomembnega terminološkega vira strokovnih prevajalcev. V članku opišemo začetno stopnjo izdelave slovenskega geslovnika dvojezičnega slovarja borznega jezika. Predstavimo dva pristopa za avtomatsko luščenje terminologije in ju na primeru luščenja iz specializiranega korpusa borznih besedil ovrednotimo.

2. Korpus borznega jezika

Korpus borznega jezika (Božinovski, 2014) je enojezični sinhroni, zaključeni, specializirani korpus. Vanj je vključenih 76 besedilnih dokumentov s področja trga kapitala v slovenskem jeziku, ki so nastala od leta 1999 dalje. Da bi korpusu, ki obsega 1.282.392 besed, zagotovili reprezentativnost oziroma uravnoteženost (Biber, 1993; Atkins idr., 1992; Arhar Holdt, 2006), smo besedila zajemali po vseh kategorijah tvorcev besedil s področja trga kapitala: zanimala so nas besedila profesorjev (ekonomistov, pravnikov) in študentov, besedila institucij trga kapitala (Ljubljanska borza, Agencija za trg vrednostnih papirjev, Centralna klirinško depotna družba, borzni člani, družbe za upravljanje,

[1] http://www.sdl.com/products/sdl-multiterm/extract.html
[2] WordSmith Tools (http://www.lexically.net/wordsmith/) v brezplačni različici omogoča omejen izpis rezultatov.
[3] SketchEngine (http://www.sketchengine.co.uk/) je v brezplačni različici na voljo 30 dni.

izdajatelji) in besedila zakonodajalca (zakonodaja) ter specializirana publicistična besedila (revija Kapital, časnik Finance itd.). Vpričo omejenih resursov smo se odločili v korpus vključiti vsa besedila, ki jih je bilo možno v razpoložljivem času pridobiti brezplačno in v primerni elektronski obliki. Izdelani korpus vsebuje znanstvene in strokovne monografije ter članke, študije in elaborate, srednje- in visokošolske učbenike, zaključna visokošolska dela, zakonodajo in predpise regulatorjev trga kapitala, brošure ter publicistična besedila. Držali smo se pravila, da se pri gradnji specializiranih korpusov zajemajo besedila v stroki uveljavljenih avtorjev (Atkins in Clear, 1992), in stremeli k čim bolj raznovrstnemu avtorstvu z namenom izključiti individualne posebnosti (Pearson, 1998). Korpus je v grobem razdeljen na tri podkorpuse: *znanstveni* vsebuje besedila uveljavljenih ekonomistov (37-odstotni delež celotnega korpusa), *strokovni* besedila borznih članov in izdajateljev ter zakonodajo in ostale predpise (42-odstotni delež), *poljudnostrokovni* pa publicistična besedila, brošure, spletne predstavitve in podobno (21-odstotni delež).[4] Večinoma so vključena celotna besedila, pri učbenikih in monografijah, ki pokrivajo področja, širša od kapitalskih trgov, pa smo poglavja, ki obravnavajo druge teme, izpustili. Vse dokumente, vključene v korpus, smo pridobili v elektronski obliki in jih tudi s pomočjo optičnih čitalnikov pretvorili v končno golo besedilo, kodirano v utf-8. Ker za vsa besedila nismo pridobili dovoljenja za javno objavo, je korpus interne narave.

3. Izbira pristopa za luščenje terminologije

K izdelavi osnutka geslovnika slovarja borznega jezika smo pristopili z avtomatskim luščenjem terminologije iz omenjenega korpusa. Po pregledu orodij za luščenje terminologije, ki podpirajo slovenščino, smo izbrali dve metodi: luščenje z orodjem WordSmith Tools, ki ga uporablja Sekcija za terminološke slovarje na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (v nadaljevanju STS SAZU), ter orodje LUIZ, ki ga v implementaciji znotraj delotoka v okolju ClowdFlows imenujemo LUIZ-CF.

3.1. WordSmith Tools (WS-L)

WordSmith Tools (različica 6, Scott, 2014) je napreden program za analizo korpusnih podatkov. Kot vstopno točko pri delu s terminologijo v specializiranem korpusu ga omenja in uporablja več avtorjev (Vintar, 2008: 92, Snoj, 2013: 2, 4, 6). Orodje smo uporabili skladno s prakso STS SAZU:[5]

V program smo uvozili celoten korpus kot golo besedilo (utf-8) ter preizkusili komponento Wordlist, ki na podlagi korpusa besedil sestavi seznam besed, urejenih po pogostosti. Gre za osnovno informacijo, ki jo potrebujemo v začetni fazi ukvarjanja s terminologijo v specializiranem korpusu (Vintar, 2008: 92). Korpus smo lematizirali z uvozom seznama lem[6] in pripadajočih besednih oblik, poleg lematizacije pa Wordlist omogoča tudi uvoz seznama "praznih besed", oz. v našem primeru seznam splošnih besed,[7] saj za razliko od običajnih seznamov praznih besed, ki vsebujejo predvsem veznike, členke, predloge itd., seznam, ki smo ga uvozili mi, zajema mnogo obširnejši nabor besed, vključujoč poleg t.i. praznih besed tudi številna lastna imena, pridevnike in druge besede splošnega jezika, s čimer se seznam bolj prilagodi terminološki nalogi.[8] Wordlist nam omogoča le luščenje enobesednih terminov,[9] v nadaljevanju pa omenjeni pristop okrajšamo z WS-L.

3.2. LUIZ-CF

Drugo orodje, ki smo ga uporabili, je LUIZ-CF, kakor imenujemo reimplementacijo sistema LUIZ (Vintar, 2010) v obliki prosto dostopnega delotoka v okolju ClowdFlows (Kranjc idr., 2012). Orodje LUIZ-CF je prosto dostopno[10] v okviru luščilnika terminologije in definicij (Pollak idr. 2012a, Pollak 2014).

Luščilnik na podlagi oblikoskladenjskih vzorcev izdela nabor kandidatov za termine, ki jih nato razvrsti glede na izračun njihove terminološke vrednosti, pri čemer primerja njihovo frekvenco v danem (specializiranem) in referenčnem korpusu[11] (Vintar, 2010). Za namen raziskave smo uporabili le del delotoka, potreben za luščenje terminologije (brez gradnikov za luščenje definicij): preko Load Corpus smo korpus naložili kot golo besedilo (utf-8), v gradniku ToTrTaLe[12] za jezikovno označevanje ter v zadnjem, ključnem gradniku Term Extraction pa smo izbrali jezik slovenščina. Kot rezultat sistema LUIZ-CF uporabnik dobi seznam terminoloških kandidatov, na katerem so eno- in večbesedni terminološki kandidati, razvrščeni na istem seznamu z

[4] Poimenovanja *znanstveni*, *strokovni* in *poljudnostrokovni* nimajo posebne metodološke vrednosti, uporabljamo jih za interno kategorizacijo besedil.

[5] Avtorica B. B. sem pristop zasnovala in izvedla skladno z opisom metodologije STS SAZU, podanim s strani dr. Tanje Fajfar, raziskovalke v STS SAZU, v osebni korespondenci dne 23. 4. in 15. 9. 2014.

[6] Uporabljeni seznam lem vsebuje 842.091 besednih oblik (100.784 lem) in je bil izdelan na podlagi leksikona besednih oblik za slovenski jezik Sloleks: http://www.slovenscina.eu/sloleks. Ker uporabljeni seznam ne vsebuje strokovnih izrazov našega področja (npr. ID, SEOnet, ETF, certifikat), opisana lematizacija pri teh kandidatih ni delovala in smo jo naknadno deloma opravili ročno.

[7] Uporabljeni seznam, ki smo ga vnesli kot seznam "praznih besed", vsebuje 343.963 besednih oblik in je bil izdelan s pomočjo Slovarja slovenskega knjižnega jezika in Slovenskega pravopisa. Oba seznama nam je zagotovil dr. A. Perdih z Inštituta za slovenski jezik Frana Ramovša ZRC SAZU.

[8] WordSmith Tools poleg osnovne funkcije Wordlist vsebuje še možnost iskanja večbesednih skupkov Clusters ter komponento Keywords, ki identificira besede v korpusu po ključnosti. Slednjih po našem vedenju na STS SAZU ne uporabljajo, zato se tudi sami osredotočimo na funkcijo Wordlist, drugi dve funkciji pa le na kratko preizkusimo. Funkcijo Keywords uporabljeni pristop delno nadomesti z zgoraj opisanim obsežnim seznamom splošnih besed. V nadaljnjem delu pa bomo v podrobnejšo primerjavo vključili tudi drugi komponenti.

[9] Osnutki geslovnikov, ki nastajajo v okviru STS SAZU, sicer vsebujejo večbesedne termine, ki jih na podlagi Wordlista preko funkcije Concord terminologi dodajo ročno.

[10] http://www.clowdflows.org/workflow/1380/

[11] V trenutni implementaciji se uporablja referenčni korpus FidaPLUS (Arhar Holdt in Gorjanc, 2007).

[12] Gradnik implementira orodje ToTrTaLe (Erjavec, 2011) in je podrobneje opisan v Pollak idr. (2012b). Izbrali smo tudi parameter za post-procesiranje, kot izhodni format pa ".txt".

normalizirano terminološko vrednostjo med 1 (najboljši) in 0 (najslabši kandidat).

3.3. Primerjava pristopov

Pristopa sta z merama natančnosti (angl. precision) in priklica (angl. recall) ovrednotena v Tabelah 1 in 2.[13] Za izračun ocene natančnosti obeh pristopov za namen sestave geslovnika smo izluščenim terminološkim kandidatom pripisali vrednosti 1 (termin) ali 0 (ni termin). V geslovnik bodo vključeni borzni termini (prim. ime stolpca *Borzni* v Tabeli 1), ovrednotili pa smo tudi termine s področij, sorodnih borznemu (korporativno pravo, računovodstvo, finance), ki sicer verjetno ne bodo vključeni v geslovnik slovarja borzne terminologije, a so z vidika terminologije vseeno zanimivi (stolpec *Vsi* tako označuje odstotek izluščenih borznih in sorodnih terminov).

Najboljši kandidati so ponavadi tisti pri vrhu seznamov. Težje pa je zajeti manj pogoste termine, zato smo za izračun ocene natančnosti pri obeh pristopih ocenili 600 kandidatov iz različnih delov seznama (imena vrstic v Tabeli 1 kažejo na zaporedno številko prvega od stotih ocenjenih kandidatov).[14]

Za izračun ocene priklica smo v naključnem dokumentu[15] iz korpusa ročno označili vse borzne termine. Tabela 2 prikaže, kolikšen odstotek tako označenih kandidatov izluščimo s posameznim orodjem ter na katerih nivojih seznamov jih najdemo (med vrhnjimi 1000, med vrhnjimi 2000 kandidati itd.).

OCENA NATANČNOSTI	WS-L		LUIZ-CF	
	Borzni	**Vsi**	**Borzni**	**Vsi**
Nivo 1	0,36	0,47	0,56	0,67
Nivo 1000	0,12	0,18	0,44	0,66
Nivo 2000	0,12	0,25	0,25	0,46
Nivo 3000	0,10	0,19	0,19	0,40
Nivo 4000	0,05	0,20	0,19	0,41
Nivo 5000	0,04	0,10	0,21	0,31
Vsi (600 kandidatov)	**0,13**	**0,23**	**0,31**	**0,48**

Tabela 1: Rezultati natančnosti luščenja (v odstotkih).

OCENA PRIKLICA	Vsi		Enobesedni	
	WS-L	**LUIZ-CF**	**WS-L**	**LUIZ-CF enobesedni**
Vrhnjih 100	0,11	0,24	0,45	0,65
Vrhnjih 1000	0,20	0,51	0,85	0,90
Vrhnjih 2000	0,23	0,70	0,95	1,00
Vrhnjih 3000		0,72		
Vrhnjih 4000		0,75		
Vrhnjih 5000		0,78		
Vrhnjih 10000		0,85		

Tabela 2: Rezultati priklica luščenja (v odstotkih).

Kot kaže Tabela 1, je pristop z orodjem LUIZ-CF bolj natančen od pristopa z orodjem Wordlist. Med vrhnjimi 100 terminološkimi kandidati je orodje LUIZ-CF namreč izluščilo več kot polovico (56 odstotkov) borznih terminov, kar je 20 odstotkov več kot pri pristopu WS-L. Tudi na nižjih nivojih je več terminov na seznamu LUIZ-CF, tako borznih kot sorodnih. Kljub temu, da natančnost na nižjih mestih večinoma pada, je med kandidati, izluščenimi z orodjem LUIZ-CF, še zmeraj približno petina borznih terminov oziroma tretjina, če upoštevamo tudi termine sorodnih področij.

Pristopa smo v nadaljevanju primerjali tudi glede priklica na podlagi ročno označenega dokumenta. Ugotovili smo, da če upoštevamo tako enobesedne kot večbesedne termine (skupaj 83 terminov), LUIZ-CF izlušči bistveno večji odstotek terminov kot WS-L. Med vrhnjimi 1000 kandidati najdemo dobro polovico terminov iz ročno označenega dokumenta (orodje WS-L izlušči le 20 odstotkov označenih terminov). Med vsemi izluščenimi kandidati, torej vključno s tistimi z zelo nizko terminološko vrednostjo, tj. do mesta 10000, je 85 odstotkov preverjanih terminov (za manjkajoče je v veliki meri kriv nepopoln nabor oblikoskladenjskih vzorcev, ki jih luščilnik zajema). Velika razlika med orodjema izhaja predvsem iz luščenja zgolj enobesednih kandidatov z orodjem WS-L v nasprotju z luščenjem tako eno- kot večbesednih z orodjem LUIZ-CF. Zaradi doslednosti primerjave in jasne utemeljitve izbire orodja za gradnjo geslovnika smo posebej izračunali še priklic za le enobesedne termine. Desni del Tabele 2 (*Enobesedni*) torej prikazuje delež izluščenih enobesednih terminov od vseh enobesednih terminov v izbranem testnem dokumentu (20 terminov). WS-L seznam tako in tako vključuje le enobesedne termine, seznam orodja LUIZ-CF pa smo skrčili na seznam enobesednih kandidatov za namen te primerjave. Na podlagi Tabele 2 smo se tudi z vidika priklica odločili za nadaljevanje sestave geslovnika z orodjem LUIZ-CF.

Za vrhnjih 100 terminoloških kandidatov vsakega orodja smo izračunali tudi stopnjo ujemanja med dvema ocenjevalcema. Poleg polstrokovnjaka prevajalca je seznam ovrednotil še področni strokovnjak ekonomist. Rezultati kažejo, da je strokovnjak kot termine označil manj kandidatov kot polstrokovnjak prevajalec, kar je bilo pričakovano in kar potrjujejo tudi izkušnje drugih (prim. Vintar, 2003; Logar Berginc idr., 2013). Kvantitativna razlika med pristopoma se je potrdila tudi pri tej oceni: na seznamu LUIZ-CF je strokovnjak kot borzne termine označil 23 odstotkov kandidatov, kot borzne in sorodne termine skupaj pa 53 odstotkov kandidatov (prim. z odstotkoma 0,56 in 0,67 pri polstrokovnjaku v Tabeli 1), medtem ko je bil pristop WS-L ocenjen bistveno slabše: potrjenih terminov je bilo 26, od tega borznih le 10 odstotkov seznama vrhnjih 100 kandidatov. Na seznamu, ki smo ga zgradili iz različnic zgornjih 100 kandidatov vsakega orodja (skupaj 140 terminov), smo izračunali splošno strinjanje (mera, ki v odstotkih izraža primere, ko sta oba ocenjevalca kandidat označila kot termin oz. netermin) in dobili rezultat 0,77. Zanimal nas je tudi koeficient kappa (Cohen, 1960), ki upošteva razliko med naključno verjetnostjo strinjanja ter opaženim strinjanjem. Za izračun kappe smo uporabili spletno orodje Vassarstats (Lowry, 2013). Kappa 0 pomeni naključno strinjanje, 1 pa popolno strinjanje. Naš rezultat je 0,5 in izraža srednjo stopnjo strinjanja (angl. moderate agreement, glej Viera in Garrett (2005)).

Da bi se dokončno prepričali o pravi izbiri orodja za nadaljevanje dela, smo na hitro preizkusili tudi drugi

[13] Ena izmed ocenjevalk v eksperimentih je strokovna prevajalka B. B., soavtorica pričujočega članka in avtorica nastajajočega borznega slovarja.

[14] Evalvacija je obsegala 600 izluščenih kandidatov na zaporednih mestih 1–100, 1000–1099, 2000–2099, 3000–3099, 4000–4099 in 5000–5099 obeh generiranih seznamov.

[15] Ljubljanska borza, d. d.: Vodnik za vlagatelje na Ljubljanski borzi, http://www.ljse.si/cgi-bin/jve.cgi?att=16774. V dokumentu z 2000 besedami je B. B. identificirala 83 terminov.

komponenti orodja WordSmith Tools, namreč funkciji Clusters in Keywords. Rezultate smo ovrednotili na naboru vrhnjih 1000 kandidatov. Funkcija za iskanje večbesednih skupkov Clusters je delovala bistveno slabše, komponenta Keywords za sezname ključnih besed pa sicer nekoliko boje od preizkušene funkcije Wordlist, vendar še vedno bistveno slabše od LUIZ-CF.

Seveda bi bilo za metodološko koherentno primerjavo samih orodij potrebno ločiti luščenje od drugih korakov (uporabiti enake načine lematizacije korpusa, upoštevati vpliv seznamov praznih besed in referenčnih korpusov itd.). Toda osnovni namen naše primerjave je bil izbrati pristop za gradnjo geslovnika strokovnega slovarja, zato smo se omejili le na primerjavo pristopov, ki so nam bili na voljo brez dodatnega dela. Na podlagi pridobljenih rezultatov smo izbrali pristop z uporabo orodja LUIZ-CF, ki tudi ne zahteva izdelave nobenih dodatnih seznamov ali predprocesiranja in je preprosto za uporabo.

4. Izdelava geslovnika in interpretacija rezultatov

S pristopom za luščenje terminologije na podlagi luščilnika LUIZ-CF smo izluščili dobrih 11000 kandidatov, od katerih smo jih ročno pregledali vrhnjih 6000. Prvih 25 terminoloških kandidatov je prikazanih v Tabeli 3. V nadaljevanju navajamo nekatera spoznanja, oblikovana med ročnim pregledovanjem izluščenih kandidatov.

Ena prvih dilem, s katero smo soočeni med ročnim pregledovanjem, je, kje in kako postaviti mejo med terminološko in splošno leksiko ter med borzno in sorodno terminologijo. Terminološka kandidata, ki ju je LUIZ-CF uvrstil na sam vrh seznama (oba imata oceno 1.0) sta *družba* in *vrednostni papir*. Medtem ko drugi ni sporen, saj gre za izrazito borzni termin, si prvega deli več strok, sega pa tudi na območje splošnega jezika; v korpusu borznega jezika so izpričani tako 1. in 2. pomen, naveden v SSKJ, ter več pravnih opredelitev družbe. Ker sestavljamo slovar borznega jezika, nas ne zanima nujno celota pomenov, zajetih v korpusu, temveč le pomeni, vezani na ožje borzno področje. Tako se bodo v geslovniku (po posvetovanju s področnim strokovnjakom) predvidoma znašli termini *borzna družba, borznoposredniška družba, delniška družba, družba za upravljanje, javna družba, investicijska družba, kapitalska družba* itd., ne pa tudi *ciljna družba, holdinška družba, hčerinska družba, družba z omejeno odgovornostjo* itd., ki spadajo na sorodno področje korporativnega prava, in to četudi imajo nekateri kandidati iz druge skupine morda višjo terminološko vrednost. Meja med termini sorodnih področji in borznimi termini je pogosto zelo tanka, saj se področje kapitalskih trgov nahaja na presečišču več strokovnih področij (prim. Logar Berginc idr., 2013, ki podobno ugotavljajo za področje odnosov z javnostmi). Po ocenah prve ocenjevalke in kot je razvidno iz zadnje vrstice Tabele 1, jih med 600 ocenjenimi kandidati za termine, ki jih je izluščilo orodje LUIZ-CF, 48 odstotkov predstavlja relevantno terminologijo: 31 odstotkov je borznih terminov in 17 odstotkov terminov s sorodnih področij (npr. *dolžniška kriza, prevzemna ponudba, bilanca stanja*), pri čemer imajo termini sorodnih področij lahko zelo visoke terminološke vrednosti. Za prepoznavanje mej med terminologijo različnih področij je potrebno termine opazovati v sobesedilu ter upoštevati mnenje stroke.

Term. vrednost	Lema	Kanonična oblika
1.000000	[družba]	<<družba>>
1.000000	[vrednosten papir]	<<vrednostni papir>>
0.671458	[trg]	<<trg>>
0.581797	[delnica]	<<delnica>>
0.399937	[člen]	<<člen>>
0.253923	[papir]	<<papir>>
0.206269	[sklad]	<<sklad>>
0.169046	[banka]	<<banka>>
0.156982	[borza]	<<borza>>
0.151592	[kapital]	<<kapital>>
0.143811	[podjetje]	<<podjetje>>
0.140058	[finančen instrument]	<<finančni instrument>>
0.127859	[nadzoren svet]	<<nadzorni svet>>
0.119758	[obveznica]	<<obveznica>>
0.109289	[odstavek]	<<odstavek>>
0.081915	[instrument]	<<instrument>>
0.070904	[trgovanje]	<<trgovanje>>
0.066481	[vrednost]	<<vrednost>>
0.066472	[tveganje]	<<tveganje>>
0.063684	[delničar]	<<delničar>>
0.057922	[član]	<<član>>
0.054564	[posel]	<<posel>>
0.054328	[naložba]	<<naložba>>
0.053878	[vlagatelj]	<<vlagatelj>>
0.052707	[obresten mera]	<<obrestna mera>>

Tabela 3: Vrhnjih 25 kandidatov, izluščenih z LUIZ-CF.

Med ročnim pregledovanjem izluščenih kandidatov smo obdržali le termine z ožjega področja kapitalskih trgov in dodali manjkajoče izraze z namenom dopolnitve posameznih pojmovnih polj ter druge ključne izraze področja. Če bi želeli v celoti slediti korpusnemu pristopu, bi morali korpus ciljno dopolnjevati z besedili, ki vsebujejo manjkajoče izraze, česar nam časovni okvir ne dopušča, poleg tega je to postopek, ki se verjetno nikoli ne konča (Atkins in Rundell 2008).

Trenutno poteka druga faza usklajevanja izrazov s področnim strokovnjakom; skupno število že potrjenih terminov za geslovnik je 1262, od katerih smo jih 91 odstotkov izluščili s pomočjo LUIZ-CF, 9 odstotkov pa jih je bilo vključenih naknadno.

Značilnost borzne terminologije, ki smo jo opazili pri ročni izdelavi geslovnika, je soobstoj številnih terminoloških variacij (npr. *prvi trgovalni dan brez upravičenja do dividend/datum brez dividend/eksdividendni datum*), kar kaže na dolgoletno odsotnost standarda. Glede rabe so korpusni podatki v nekaterih primerih v popolnem nasprotju s prepričanjem stroke (stroka denimo zadnja leta zagovarja rabo termina *finančni instrument* namesto *vrednostni papir*, četudi je slednji bistveno pogostejši v praktično vseh večbesednih zvezah, npr. *dolžniški vrednostni papir* se v korpusu pojavi 247-krat, *dolžniški finančni instrument* pa 12-krat). Če želi nastajajoči borzni slovar prevzeti normativno vlogo, bo zato potrebno soočiti rabo, kot se je z leti ustalila, in mnenje stroke.

Seznam izluščenih terminoloških kandidatov nam je sprva služil tudi kot osnova za dopolnitev terminološke

zbirke s terminološkimi kolokacijami.[16] Vendar se je upoštevanje kolokacij na začetni stopnji sestave geslovnika izkazalo za preveč zamudno, saj je zaradi velike količine pojmovno nerazvrščenih kandidatov oteževalo delo. Zato smo jih na tej stopnji izključili iz obravnave in se bomo k njim vrnili v fazi izdelave geselskih člankov.

Podkorpus PS	Podkorpus S	Podkorpus Z
delnica	družba	trg
milijon evrov	vrednostni papir	vrednostni papir
vrednostni papir	član	podjetje
obrestna mera	nadzorni svet	kapital
evro	finančni instrument	delnica
trg	odstavek	banka
odstotek	delnica	papir
milijarda evrov	borznoposredniška družba	obrestna mera
ljubljanska borza	papir	obveznica
obveznica	član	naložba

Tabela 4: Vrhnjih 10 kandidatov po podkorpusih (poljudnostrokovni, strokovni, znanstveni).

Dodatno se nam je zdelo zanimivo preučiti rezultate luščenja terminologije za posamezne podkorpuse, torej za podkorpus znanstvenih besedil, podkorpus strokovnih besedil in podkorpus poljudnostrokovnih besedil. Primerjali smo vrhnjih 100 izluščenih kandidatov vsakega podkorpusa, po prvih 10 predstavljamo v Tabeli 4. Analiza terminov kaže, da so besedila po strukturi besedišča v posameznih podkorpusih zelo raznolika. Največja opazna odstopanja so pravno-zakonodajni izrazi (*člen, odstavek, določba, zakon*) v podkorpusu strokovnih besedil, v katerem prevladujejo predpisi s področja trga kapitala, ter izrazi za vrednosti (*milijon/milijarda evrov/dolarjev*) v podkorpusu poljudnostrokovnih besedil. Največji delež pravilno izluščenih terminov je sicer luščilnik prepoznal v znanstvenem podkorpusu (79 %).

5. Predlogi za izboljšavo luščilnika

V tem razdelku analiziramo pomanjkljivosti luščenja terminologije z LUIZ-CF z vidika uporabnosti za sestavo geslovnika strokovnega področja.

Obravnava terminoloških variacij. Za borzno terminologijo je v odsotnosti standarda značilna variantnost, saj raba ni ustaljena. LUIZ-CF variacije prikaže kot različne kandidate, ki jih je treba ročno iskati in združevati. Gre tako za pisne variacije (npr. finančna institucija vs. finančna inštitucija, SEOnet vs. SEO-net vs. seo-net vs. SEO net) kot oblikoslovne (posoja/posojanje vrednostnih papirjev) in skladenjske variacije (kapitalski trg vs. trg kapitala).

S tem je tesno povezana *obravnava enakovrednih poimenovanj.* V izluščeni terminologiji je mnogo primerov, ko za en pojem obstajata dve različni enakovredni poimenovanji, ki jih LUIZ-CF prikaže neodvisno drugo od drugega. Gre za pare, za katere bi bilo z vidika pojmovnega pristopa zaželeno, da bi jih orodje ponudilo skupaj. Glavni podkategoriji sta domač izraz – tujka (npr. vlagatelj/investitor, izračun/kliring, navzkrižje/konflikt interesov) in razvezan izraz – kratica (celotna globina trga/CGT, vzajemni sklad/VS).

Lastna imena LUIZ-CF enakovredno razvršča med kandidate za termine, vključno z osebnimi imeni (priimki: Berk, Simoneti; imena indeksov: Eurostoxx, Dow Jones, NASDAQ) in imeni institucij (Ljubljanska borza, Wall Street, Ameriška centralna banka), pri katerih je odločitev za ali proti vključitvi v slovar zahtevna. Koristen bi bil parameter, s katerim bi lahko seznama ločili.

Sorodni termini. Kot že večkrat omenjeno zgoraj, je težava tovrstnega luščenja tudi nehoten zajem številnih kandidatov s sorodnih področij (nekaj primerov parov borznih/neborznih terminov: depozitar/depozit, trgovalni račun/tekoči račun, delniška družba/komanditna družba, devizna izmenjava/devizni tečaj). Pri nadgradnji LUIZ-CF bi bilo v ta namen smiselno preizkusili strojne metode aktivnega učenja (angl. active learning).

Terminološke kolokacije. V fazi izdelave geslovnika so problematične tudi terminološke kolokacije, saj se izkaže, da na začetni stopnji gradnje slovarja zmanjšajo preglednost zbranega gradiva in upočasnjujejo inventarizacijo terminologije. LUIZ-CF jih je izluščil zelo veliko (primeri parov borzni termin/terminološka kolokacija: osnovni kapital/povečanje osnovnega kapitala, skupščina delničarjev/sklic skupščine delničarjev, avkcija/trg v avkciji). V kolikor pride do odločitve za vključitev kolokacij v terminološki slovar, so te potrebne šele kasneje in ne že v fazi sestave geslovnika, zato bi bilo smiselno, da so na seznamu izluščenih kandidatov za termine prikazane ločeno.

Omenili smo že, da priklic ni bil 100 odstoten zaradi nekaterih nepokritih *oblikoskladenjskih vzorcev*.

6. Zaključki in nadaljnje delo

V članku smo predstavili poskus luščenja terminologije iz korpusa borznega jezika, ki ga sestavljajo besedila znanstvenega, strokovnega in poljudnostrokovnega značaja. Terminologijo smo luščili z uporabo luščilnika LUIZ (Vintar, 2010) v spletni implementaciji LUIZ-CF (Pollak idr. 2012a, Pollak 2014), ter njegovo natančnost in priklic primerjali s podobnim pristopom z orodjem Wordlist (WordSmith Tools). Za temeljito primerjavo orodij bi morali pri obeh orodjih ločiti luščenje od predprocesiranja ter v skupku orodij WordSmith Tools natančneje preizkusiti še funkciji Keywords in Clusters.

Na podlagi rezultatov smo za izdelavo nastajajočega geslovnika slovarja borznega jezika izbrali orodje LUIZ-CF, terminologijo pa v sodelovanju s področnim strokovnjakom dopolnjujemo ročno. Analizirali smo lastnosti izluščenih terminoloških kandidatov in identificirali mesta za možne izboljšave luščilnika (obravnava terminoloških variacij, terminov sorodnih področij, kolokacij). Vzporedno izboljšujemo tudi sam delotok, da bi luščenje terminologije potekalo hitreje. V

[16] Izbirali smo jih izrazito selektivno, saj je namen njihove vključitve točno določen in dvojen, prilagojen zlasti uporabniku nestrokovnjaku: po eni strani naj bi z dodatnimi informacijami o pojmovnem polju termina okrepile razumevanje njegove razlage (Bergenholtz in Tarp, 1995) in uporabniku ponudile leksikalno okolje termina (Atkins in Rundell, 2008), po drugi strani pa podprle prevajalsko funkcijo nastajajočega slovarja (za ta namen bodo izbrane kontrastivno zanimive, torej netransparentne kolokacije). V zbirko kljub njihovi terminološkosti nismo želeli vključiti vseh kolokacij s področja kapitalskih trgov, temveč le tiste, ki podpirajo omenjena cilja.

nadaljevanju raziskave bomo delotok, predstavljen v Pollak (2014), uporabili za luščenje definicij iz omenjenega korpusa ter ovrednotili njegovo uporabnost.

Zahvala

Zahvaljujemo se dr. Tanji Fajfar, dr. Mojci Žagar Karer in dr. Andreju Perdrihu z Inštituta za slovenski jezik ZRC SAZU, ki so drugi avtorici članka prijazno pomagali in ji posredovali informacije o svojem pristopu k uporabi orodja WordSmith Tools ter ji omogočili uporabo svojih seznamov lem in praznih besed.

7. Literatura

Ahmad, K., L. Gillam, in L. Tostevin, 2007. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Rerieval (WILDER). *Proceedings of the Eight Text REtrieval Conference (TREC-8)*: 717–724.

Arhar Holdt, Š., 2006. Gradnja specializiranega korpusa. *Jezik in slovstvo,* 51(1): 53–67.

Arhar Holdt, Š., in V. Gorjanc, 2007. Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52 (2): 95–110

Atkins, S., in J. Clear, 1992. Corpus design criteria. *Literary and Linguistic Computing*, 7(1): 1–16.

Atkins, B.T.S, J. Clear, in N. Ostler, 1992. Corpus design criteria. *Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing* 7(1): 1–16.

Atkins, B.T.S, in M. Rundell, 2008. *The Oxford guide to practical lexicography*. Oxford/New York: Oxford University Press.

Bergenholtz, H., in S. Tarp (ur.), 1995. *Manual of Specialised Lexicography*. Amsterdam/Philadelphia: Benjamins Publishing Company.

Biber, D., 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing,* 8(4): 243–257.

Božinovski, B. 2014 (v pripravi). *Problematika slovensko-angleškega strokovnega izrazoslovja s področja borznega poslovanja*. Doktorska disertacija, Univerza v Ljubljani.

Cohen, J., 1960. A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37–46.

Čas, M., in T. Rotar, 1994. *Borzni izrazi: s trojezičnim slovarjem*. Maribor: Kapital.

Erjavec, T., 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. *Proceedings of the ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL 2011)*.

Frantzi, K. T., in S. Ananiadou, 1999. The CValue/NCValue domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3): 145–179.

Kozakov, L., Y. Park, T. Fin, Y. Drissi, Y. Doganata, in T. Cofino, 2004. Glossary extraction and utilization in the information search and delivery system for IBM technical support. *IBM Systems Journal*, 43(3): 546–563.

Kranjc, J., V. Podpečan, in N. Lavrač, 2012. ClowdFlows: A cloud based scientific workflow platform. *Proceedings of ECML/PKDD-2012 (2)*, Springer LNCS 7524: 816–819.

Lefever, E., M. Lieve, in V. Hoste, 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. *Proceedings of the 12th Conference of the European Chapter of the ACL*: 496–504.

Logar Berginc, N., Š. Vintar, in Š. Arhar Holdt, 2013. Terminologija odnosov z javnostmi: korpus - luščenje - terminološka podatkovna zbirka. *Slovenščina 2.0*, 1(2): 113–138.

Lowry, R.. 2013. *Kappa as a measure of concordance in categorical sorting.* http://vassarstats.net/kappa.html. Zadnji dostop: 19. september, 2014.

Macken, L., E. Lefever, in V. Hoste, 2013. TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology,* 19(1): 1–30.

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

Pollak, S., A. Vavpetič, J. Kranjc, N. Lavrač, in Š. Vintar, 2012a. NLP workflow for online definition extraction from English and Slovene text corpora. *Proceedings of the 11th Conference on Natural Language Processing*, 53–60.

Pollak, S., N. Trdin, A. Vavpetič, in T. Erjavec, 2012b. NLP Web Services for Slovene and English: Morphosyntactic tagging, lemmatisation and definition extraction. *Informatica,* 36: 441–449.

Pollak, S., 2014. *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov.* Doktorska disertacija, Univerza v Ljubljani.

Sclano, F., in P. Velardi, 2007. TermExtractor: a Web application to learn the common terminology of interest groups and research communities. *Proceedings of the 9th Conference on Terminology and Artificial Intelligence TIA 2007*: 8–9.

Scott, M., 2014a. *WordSmith Tools version 6* (and *WordSmith Tools Manual*), Liverpool: Lexical Analysis Software.

Snoj, M., 2013. Zaključno poročilo raziskovalnega projekta – 2013. Oznaka poročila: ARRS-RPROJ-ZP-2013/203.

Viera, A. J., in J. M. Garrett, 2005. Understanding interobserver agreement: The Kappa Statistic. *Family Medicine*, 37(5): 360–363.

Vintar, Š., 2003. *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov*. Doktorska disertacija, Univerza v Ljubljani.

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

Vintar, Š., 2010. Bilingual term recognition revisited. The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2): 141–158.