Luščenje terminologije iz angleško-slovenskih vzporednih in primerljivih korpusov

Špela Vintar
Oddelek za prevajalstvo, Filozofska fakulteta Univerze v Ljubljani

Abstract

The paper describes LUIZ, a bilingual term recognition system that has been developed for the Slovene-English language pair. The system is a hybrid term extractor using morphosyntactic patterns and statistical ranking to propose domain-specific expressions for each of the two languages, whereupon translation equivalents between the languages are identified using the innovative bag-of-equivalents approach. This simple but effective method is based on the Twente word aligner to obtain a lexicon of single word translation pairs and their probability scores, which is then used to identify correspondences between multi-word terms.

The bilingual term recognition system has been tested and evaluated on three parallel subcorpora from the tourism, accounting and military domain. Average precision of the term alignment component is 0.83, whereby only fully equivalent and domain-relevant terms were counted as positives. Another advantage of the described approach is the fact that we successfully detect term variants and multiple translations of a candidate multi-word term. Since our term alignment method does not require sentence-aligned corpora it can be used with comparable corpora, provided we already have a domain-specific lexicon or dictionary of single-word correspondences. The paper concludes with some thoughts on the users of term recognition systems and their needs based on our observations from the online version of the system.

Ključne besede: dvojezično luščenje terminologije, evalvacija luščenja terminologije, poravnava terminov, vzporedni korpusi, primerljivi korpusi

1 UVOD

Samodejno prepoznavanje ali luščenje terminološko relevantnih leksikalnih enot (angl. *automatic term recognition* ali *term extraction*) je raziskovalno področje v sklopu računalniškega in korpusnega jezikoslovja, ki je v zadnjih dveh desetletjih doživljalo živahen razvoj in katerega glavni namen je identifikacija eno- in večbesednih področnih terminov v specializiranem korpusu, in to brez ali z minimalno človekovo pomočjo. Sistemi za samodejno luščenje terminologije so danes na voljo za številne jezike in jezikovne pare, del raziskovalnih naporov na tem področju pa je namenjen tudi njihovi evalvaciji. S pojavitvijo tržnih proizvodov, ki ponujajo luščenje terminologije za kateri koli jezik, in z vse boljšo pokritostjo različnih jezikov s temeljnimi orodji za jezikoslovno analizo se v zadnjih letih zdi, da je ta jezikovnotehnološki problem v veliki meri razvozlan, čeprav podrobnejši pogled v uspešnost teh sistemov in uporabnost rezultatov razkriva še mnogo priložnosti za izboljšave.

Pričujoči prispevek predstavlja področje samodejnega pridobivanja terminoloških izrazov iz eno- in večjezičnih specializiranih korpusov, v okviru tega pa predvsem zgradbo in evalvacijo dvojezičnega luščilnika terminologije za angleško-slovenska besedila LUIZ, ki smo ga razvili že leta 2004 in odtlej uporabili v številnih projektih in na zelo raznolikih strokovnih področjih. S tem smo pridobili dragocene povratne informacije o potrebah različnih uporabnikov terminologije, posebnostih strokovnih področij in slabostih samega sistema. Od leta 2008 je poskusna različica luščilnika za slovenščino na voljo tudi kot spletna aplikacija, s čimer se je krog uporabnikov in vir odzivov še razširil. V nadaljevanju v drugem razdelku podajamo pregled pomembnejših metod luščenja tako v eno- kot v dvojezičnem kontekstu, nato pa v tretjem razdelku opišemo sistem LUIZ, ki vključuje izviren način iskanja prevodnih ustreznic z »vrečo ustreznic«. V četrtem razdelku predstavimo evalvacijo dvojezične poravnave terminov, katere natančnost v povprečju znaša okrog 0,84. V zadnjem razdelku razpravljamo o tipičnih uporabnikih sistemov za luščenje terminologije, ki jih glede na njihove specifične potrebe razdelimo na tri kategorije, prispevek pa sklenemo z vizijo o korpusno-terminoloških tehnologijah prihodnosti.

2 PREGLED METOD ZA SAMODEJNO LUŠČENJE IZRAZJA

V zadnjih dveh desetletjih smo bili na področju samodejnega luščenja terminologije priča izredno živahni raziskovalni dejavnosti. Večina tradicionalnih pristopov k luščenju se opira bodisi na porazdelitvene lastnosti terminov, kar pomeni, da merijo njihovo pogostost v specializiranem korpusu ali zbirki dokumentov

ter jo primerjajo s pogostostjo v splošnem (referenčnem) korpusu (Ahmad et al. 1992, Ananiadou 1994), bodisi uporablja oblikoskladenjske vzorce za zajem terminologije na podlagi njihove oblike. Večina zgodnjih pristopov pravzaprav uporablja kombinacijo obeh tehnik, in sicer se s pomočjo besednovrstnih vzorcev najprej izlušči začetni seznam potencialnih leksikalnih enot, nato pa se uporabi sito »terminološkosti«, za kar različni avtorji predlagajo različne numerične metode (Bourigault et al. 1996, Heid 1998, Mima in Ananiadou 2000, Nakagawa 2000, Uchimoto 2000, glej tudi Kageura et al. 2000 za pregled pristopov, predstavljenih na delavnici NTCIR-1). Nekateri sistemi pri tem namesto oblikoskladenjskih vzorcev uporabljajo polno skladenjsko razčlembo, kar se še posebej obnese pri jezikih z manjšo oblikoslovno razvejanostjo, kot je angleščina (Bernth et al. 2003).

Inovativnejši pristopi k luščenju presegajo zgolj kombinacijo statističnih in jezikoslovnih lastnosti terminov in vključujejo semantične informacije; tu gre predvsem za navezavo luščenja terminologije na samodejno gradnjo ontologij in tehnologije znanja. Številni avtorji tako uporabljajo metode rudarjenja besedil in skušajo odkrivati tudi semantična razmerja med pojmi (Collier et al. 2001; Nenadić et al. 2002; Mima et al. 2006). Druga veja raziskav, ki izvira predvsem iz francosko govorečih držav, področje luščenja terminologije razširi s sistematično obravnavo terminoloških variacij, ki lahko v določenih pogojih tudi pripomorejo pri sami identifikaciji terminološko relevantnih zvez (Jacquemin 2001; Daille 2003). Tiedemann (2001) predlaga metodo, pri kateri ugotavljanje terminološkosti v enem jeziku poteka s pomočjo vzporednega korpusa; dvojezična poravnava terminov namreč lahko služi kot merilo za stabilnost terminološke zveze. Na soroden način Oh et al. (2000) uporabljajo strojno prevajanje (glej tudi Kageura et al. 2004).

Področje dvojezičnega luščenja terminologije je nekoliko manj raziskano, večina pristopov pa to nalogo razstavi na enojezično luščenje za vsak jezik posebej, čemur sledi postopek iskanja prevodnih ustreznic med izluščenimi kandidati. Za dvojezično luščenje so najbolj primerni vzporedni korpusi, pri katerih uporabljamo statistične metode za ugotavljanje terminološke ekvivalence med jeziki. Zgodnje raziskave se ukvarjajo zgolj s poravnavo enobesednih enot (Hiemstra 1998; Melamed 2000), Ahrenberg et al. (1998) pa vključujejo tudi večbesedne enote. Kwong et al. (2004) opisujejo dvojezično luščenje terminologije iz kitajsko-angleškega vzporednega korpusa pravnih besedil, pri tem pa za iskanje ustreznic uporabljajo primerjavo pogostostnih porazdelitev; metoda dosega 79-odstotno natančnost. Izvirno in uspešno metodo predstavlja tudi Gaussier (1998), ki za iskanje francosko-angleških eno- in večbesednih terminoloških kandidatov predlaga na grafih temelječ mrežni model in za prvih 500 kandidatov dosega 90-odstotno natančnost.

Ker je vzporedne korpuse za nekatera področja in jezikovne pare težko in zamudno zagotoviti, se številne raziskave ukvarjajo z dvojezičnim luščenjem iz nevzporednih korpusov. Mann in Yarowsky (2000) poročata o metodi, ki prevodno ustreznost ugotavlja s pomočjo sorodnic (*cognates*).[1] Tako gradita dvojezične leksikone iz primerljivih korpusov za poljubni jezikovni par. Vzporedno s tem so se pričeli razvijati tudi numerično kompleksnejši pristopi, denimo Fung in McKeown (1997), ki za ugotavljanje prevodnih ustreznic uporabljata kontekstne vektorje. Njun algoritem temelji na seznamu znanih parov prevodnih ustreznic, ki služijo za »seme«, nato pa se izračunavajo matrike podobnosti med vsako besedo in semensko besedo. Na podlagi teh vektorjev sopojavljanja je mogoče izračunati prevodno ustreznost, pri čemer je povprečna natančnost za prvo predlagano ustreznico okrog 30 %. Gaussier et al. (2004) nadaljujejo v podobni smeri in opisujejo metodo za dvojezično luščenje terminologije iz primerljivih korpusov s pomočjo latentne semantične analize, pri tem pa se za prevod kontekstnega vektorja uporablja splošni dvojezični slovar. Povprečna natančnost pri njihovem pristopu dosega že 44 %.

3 LUIZ – DVOJEZIČNI LUŠČILNIK IZRAZJA ZA ANGLEŠKO-SLOVENSKI JEZIKOVNI PAR

Slovenščina je oblikoslovno izredno bogat jezik, zato je pri večini jezikovnotehnoloških metod lematizacija nujna stopnja predobdelave, saj šele statistika lem prikaže realistično podobo pogostostnih razmerij v korpusu. Po drugi strani so večbesedne terminološke enote, ki jih želimo izluščiti, sestavljene iz besednih oblik, med katerimi vladajo pomembna ujemalna razmerja. Pri postopku luščenja moramo tako najti občutljivo ravnovesje med normalizacijo slovničnih kategorij in njihovim ohranjanjem.

Sistem LUIZ smo razvili leta 2003 v dveh različicah. Statistični luščilnik je temeljil na vhodnih podatkih v obliki neoznačenih poravnanih besedil, hibridna različica pa je uporabljala oblikoskladenjsko označena in lematizirana besedila ter spisek oblikoskladenjskih vzorcev za luščenje. Po izvedbi prvih evalvacijskih preskusov (Vintar 2003), katerih rezultati niso bili preveč obetavni, ter po objavi prvega brezplačnega označevalnika in lematizatorja za slovenščino (Erjavec et al. 2005) smo nadaljnji razvoj statistične različice opustili.

Sedanja različica sistema deluje kot hibridni dvojezični luščilnik terminologije, ki kot vhodne podatke pričakuje vzporedni ali primerljivi korpus, vrne pa eno – in dvojezični seznam terminoloških kandidatov. Zgradbo sistema kaže Slika 1.

[1] Sorodnice (angl. *cognates*) so na področju računalniškega jezikoslovja besede, ki so – navadno zaradi skupnega izvora – v dveh ali več jezikih enake ali podobne. Sem sodijo tako internacionalizmi (*taxi, hotel, pizza*) kot lastnoimenske enote (*London, George Bush, Avstrija*).

3.1 Postopek luščenja

Luščenje izrazja poteka ločeno za vsakega od obeh jezikov, pri čemer so korpusna besedila lematizirana in oblikoskladenjsko označena. Za vsakega od obeh jezikov uporabljamo seznam terminološko relevantnih oblikoskladenjskih vzorcev, ki zajema predvsem samostalniške besedne zveze dolžine do pet besed. Za angleščino so ti vzorci pravzaprav zaporedja besednih vrst (npr. samostalnik + samostalnik, pridevnik + samostalnik), za slovenščino pa uporabljamo tudi kategoriji sklona in števila, s čimer dosežemo boljše ločevanje med sosednjimi samostalniškimi zvezami (npr. P---ei S---ei, kar pomeni zaporedje pridevnika v imenovalniku ednine ter samostalnika v imenovalniku ednine). V nadaljevanju opisani poskusi temeljijo na seznamih 14 slovenskih in 16 angleških oblikoskladenjskih vzorcev, pri čemer je seznam vzorcev mogoče spremeniti v skladu s specifičnimi zahtevami uporabnika luščilnika.

Sistem iz korpusa najprej izlušči vse besedne zveze, ki ustrezajo enemu od določenih vzorcev, nato pa jih razvrsti glede na terminološkost. Terminološkost (W) izluščene besedne zveze a, ki vsebuje n besed, se izračuna po naslednji formuli:

$$W(a) = \frac{f_a^2}{n} \cdot \Sigma \left(\log \frac{f_{n,D}}{N_D} - \log \frac{f_{n,R}}{N_R} \right)$$

kjer je f_a absolutna pogostost besedne zveze v specializiranem korpusu, $f_{n,D}$ in $f_{n,R}$ sta pogostosti vsake posamezne vsebovane besede v specializiranem in referenčnem korpusu, N_D in N_R pa sta velikosti obeh korpusov v pojavnicah.

Osnovna ideja izračuna terminološkosti je predpostavka, da večbesedne terminološke enote sestavljajo besede, ki so tudi same terminološko pomembne, merilo terminološke pomembnosti pa je primerjava med pogostostjo besede v specializiranem in splošnem/referenčnem korpusu. Če tako denimo primerjamo terminološkost enot a – *armored personnel carrier* in b – *rapid change*, ki se v korpusu vojaških besedil obe pojavljata dvakrat, nam primerjava s pogostostmi iz korpusa BNC daje naslednji vrednosti W:

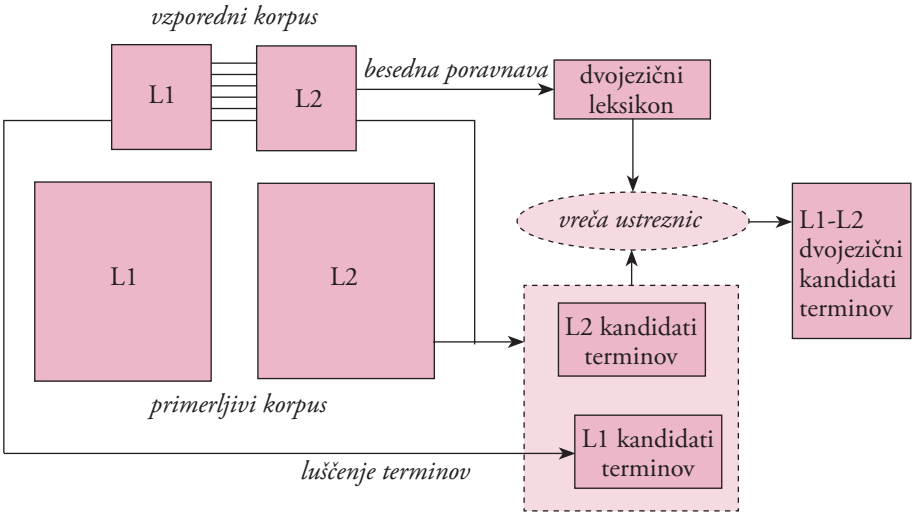
$$W(a) = 2^2/3 * (3{,}73 + 1{,}77 + (-0{,}13)) = 5{,}32$$

$$W(b) = 2^2/2 * (0{,}9 + 0{,}39) = 2{,}58$$

Kot splošnojezikovni korpus uporabljamo za slovenščino FidoPlus[2] in za angleščino BNC.[3]

[2] http://www.fidaplus.net

[3] Uporabljeni so bili prosto dostopni besedni seznami iz korpusa BNC, ki jih je objavil Mike Scott na strani http://www.lexically.net/downloads/BNC_wordlists/.



Slika 1: Zgradba sistema LUIZ

3.2 Ugotavljanje prevodne ustreznosti – metoda »vreče ustreznic«

Sprva sistem izlušči terminološke kandidate za vsak jezik posebej, v naslednjem koraku pa želimo izluščene besedne zveze povezati v pare prevodnih ustreznic. Osnova za ugotavljanje prevodne ustreznosti je statistična besedna poravnava (*word alignment*) s prosto dostopnim programom Twente, ki iz vzporednega korpusa za vsako besedo izračuna statistično najverjetnejše prevodne ustreznice (Hiemstra 1998). Program Twente uporablja algoritem EM za izračun prevodnih verjetnosti v dveh simetričnih modelih besedne poravnave, pri čemer model A privzema, da se vsaka beseda v izvirnem stavku prevede v eno samo besedo v ciljnem stavku, za izravnavo razlik v dolžini stavkov pa se uvede še »prazna« beseda (null). Model B dopolnjuje prvega s tem, da dopušča tudi besedna ustrezanja ena-na-več in več-na-ena.

Čeprav je za besedno poravnavo na voljo nekaj bolj razširjenih prosto dostopnih orodij, še posebej Uplug in Giza++ (Tiedemann 2003), je za pristop z vrečo ustreznic ključnega pomena, da je izhodiščna beseda lahko poravnana z več možnimi prevodnimi ustreznicami ter da se ustreznice predlagajo tudi za besede z nizko pogostostjo pojavitve v korpusu.

Še pred besedno poravnavo iz vseh besedil odstranimo prazne besede ter besedne oblike pretvorimo v leme. Rezultat besedne poravnave je dvojezični leksikon, ki

za vsako besedno lemo v korpusu podaja niz ustreznic z njihovimi verjetnostmi. Po končani izdelavi dvojezičnega leksikona lahko pričnemo s poravnavo večbesednih terminov, ki smo jih izluščili iz nevzporednega korpusa, pri čemer nam metoda vreče ustreznic omogoča izbiro najboljše prevodne ustreznice za izvirni večbesedni termin. Če denimo iščemo slovensko ustreznico za vojaški termin *destruction of anti-personnel mines*, dvojezični leksikon vsebuje naslednje vnose:

destruction	*uničevanje*	*0.86*	*uničenje 0.14*
anti-personnel	*protipehoten*	*1.00*	
mine	*mina*	*1.00*	

Vse štiri predlagane slovenske besede zberemo v »vrečo«, nato pa med izluščenimi slovenskimi termini poiščemo tistega, ki se jim najbolje prilega, in sicer tako, da je mera ustrezanja preprosto vsota vseh posamičnih verjetnosti deljena s številom besed v slovenskem terminu. Za izbrani angleški izraz tako dobimo štiri prevodne ustreznice, od katerih sta pravilni dve:

destruction of anti-personnel mines	***uničevanje protipehotnih min***	***0.95***
	uničenje protipehotnih min	***0.71***
	uporaba protipehotnih min	*0.66*
	prepoved protipehotnih min	*0.66*

Opisani pristop ima dve prednosti. Prvič nam omogoča, da za izbrani termin v izvirniku poiščemo več ustreznic, kar je še posebej dragoceno pri strokovnih področjih z manj ustaljeno terminologijo in visoko variabilnostjo v izrazju. Iz zgornjega primera je denimo razvidno, da sta tako uničevanje protipehotnih min kot uničenje protipehotnih min možna prevedka izvirnega termina in predstavljata terminološko variacijo. Drugič pa je s tem pristopom mogoče najti ustreznice tudi za termine z besedami, za katere nam dvojezični leksikon predlaga napačne ali nepopolne prevode, kot je razvidno iz spodnjega primera za izraz *early warning system (sistem za zgodnje opozarjanje)*:

early	*(null) 0.28*	*zgodnji 0.20*	*opozarjanje 0.20*	*prej 0.20*	...
warning	*opozorilen 0.40*	*grožnja 0.20*	*zgodnji 0.20*	*opozarjanje 0.20*	
system	*sistem 0.97*	*sistemski 0.01*	*(null) 0.01*		

4 EVALVACIJA SISTEMA LUIZ

Evalvacija sistemov za samodejno luščenje terminologije je izredno kompleksna naloga, zato tudi ni enotne metodologije vrednotenja rezultatov, ki bi bila pravična do vseh sistemov, uporabnikov in namenov luščenja. Običajni način evalvacije

jezikovnotehnoloških sistemov z merjenjem natančnosti, priklica in vrednosti F tu ni najbolj primeren, saj za večino specializiranih korpusov, iz katerih samodejno luščimo izrazje, ne poznamo natančnega števila vsebovanih terminov in ga tudi ne moremo enostavno določiti (Vivaldi in Rodriguez 2007). Poleg tega je razlikovanje med termini in netermini vse prej kot enostavno, saj se o terminološkosti – kot kažejo eksperimenti – tudi strokovnjaki med seboj težko sporazumejo (Estopà Bagot 1999).

Slabost tradicionalnih pristopov k evalvaciji je tudi njihova binarnost v smislu, da je terminološke kandidate vselej možno označiti bodisi kot termin ali netermin, čeprav bi jih po intuiciji morda laže razvrščali po večstopenjski lestvici; nenazadnje to predlaga tudi večina teoretikov terminološke vede.

Pričujoči prispevek se v prvi vrsti posveča evalvaciji dvojezične poravnave terminov pri sistemu LUIZ, saj je bila kakovost samega luščenja terminov že ovrednotena v sklopu različnih prejšnjih eksperimentov (Vintar 2003, Vintar 2004, Vintar 2009). Pri prvi omenjeni evalvaciji smo terminološke kandidate vrednotili s pomočjo strokovnjakov, ki so za ocenjevanje uporabljali petstopenjsko lestvico s kategorijami *je termin, je za stroko specifični izraz, vsebuje termin* itd. Kljub nizki stopnji soglašanja med obema strokovnjakoma in težavni pretvorbi opisnih oznak v enotno številsko mero je bilo v povprečju 49 % terminoloških kandidatov označenih bodisi kot termin ali za stroko pomemben izraz. V poznejših evalvacijskih eksperimentih, ki smo jih izvajali na področjih informacijske tehnologije, jedrske tehnike in računovodstva, smo uporabljali binarno razvrščanje, natančnost luščenja za slovenščino pa se je gibala med 0,65 in 0,83.

V dvojezičnem kontekstu je kakovost luščenja sestavljena iz treh delov, in sicer terminološkosti kandidatov v prvem in drugem jeziku ter prevodne ustreznosti med njima. V nadaljevanju opisujemo evalvacijo modela za poravnavo izluščenih terminov pri sistemu LUIZ po metodi vreče ustreznic, in sicer na treh strokovnih področjih.

4.1 Področja in korpusi

Za potrebe evalvacijskega eksperimenta smo uporabili vzporedne slovensko-angleške korpuse s področij turizma, računovodstva in vojaštva. Korpusi so vsebovali naslednje besedilno gradivo:

- turizem: 130.000 pojavnic, Strategija razvoja turizma v Sloveniji 2007-2011
- računovodstvo: 280.000 pojavnic, Slovenski računovodski standard I in II

- vojaštvo: 110.000 pojavnic, Strateški pregled obrambe RS ter obvestila za javnost Ministrstva za obrambo RS

Pri vseh podkorpusih smo izvedli stavčno poravnavo, tokenizacijo, lematizacijo, oblikoskladenjsko označevanje s ToTaLe (Erjavec et al. 2005) ter pretvorbo v enotni zapis XML v kodnem naboru UTF-8. Za obdelavo z besednim poravnavalnikom Twente smo iz besedil odstranili prazne besede in besedne oblike pretvorili v leme, stavčno poravnavo pa ohranili. Tako smo za vsako področje pridobili dvojezični verjetnostni leksikon enobesednih enot, in sicer v obe smeri (slovensko-angleški in angleško-slovenski).

Luščenje terminoloških kandidatov poteka za vsak jezik posebej. Terminološkost enobesednih enot izračunamo na podlagi primerjave relativne pogostosti besede v specializiranem in referenčnem korpusu, nato iz korpusa s pomočjo oblikoskladenjskih vzorcev izluščimo večbesedne enote in vsaki enoti izračunamo terminološkost po prej navedeni enačbi. Tabela 1 vsebuje podatke o velikosti posameznih podkorpusov in številu izluščenih terminoloških kandidatov.

	Turizem		Računovodstvo		Vojska	
	sl	an	sl	an	sl	an
Velikost korpusa	62,481	72,123	118,650	161,832	49,795	59,509
Št. izluščenih	2,152	1,772	3,194	2,520	1,803	1,421

Tabela 1: Velikosti korpusov in število izluščenih terminov

4.2 Poravnava terminov in evalvacija

V naslednjem koraku želimo za vsakega terminološkega kandidata v enem jeziku poiskati eno ali več prevodnih ustreznic v ciljnem jeziku. Ker nam orodje Twente izdela dvojezični leksikon v obe smeri, poravnavo terminov prav tako izvajamo iz slovenščine v angleščino in obratno, saj nas zanimajo morebitne razlike v natančnosti. Sistem tako za vsako enoto poišče možne prevodne ustreznice v lematizirani in kanonični obliki ter izračuna stopnjo ustrezanja, nato pa obdržimo le prvo in drugo najboljšo ustreznico – slednja je namreč pogosto variantni prevod izvirnega termina.

Pri evalvaciji smo uporabili prvih 300 parov terminov, razvrščenih glede na stopnjo ustrezanja. Pri ocenjevanju prevodne ustreznosti smo uporabili stroga merila, kar pomeni, da smo za pravilne šteli le primere, pri katerih je bil ciljni termin popolna in pravilna prevodna ustreznica izvirnega termina. Natančnost za posamezna področja ter za obe jezikovni smeri povzema Tabela 2.

	Turizem	Računovodstvo	Vojska	Povprečno
sl-an	0.636	0.846	0.970	0.817
an-sl	0.832	0.836	0.880	0.849

Tabela 2: Natančnost poravnave terminov

Zagotovo lahko trdimo, da so zgornji rezultati spodbudni in kažejo, da je sistem v povprečju za prek 80 odstotkov terminov predlagal pravilen prevod, pri tem pa niti smer prevajanja niti velikost vzporednega korpusa ne igrata bistvene vloge. Pri turističnem korpusu je bila za slovensko-angleški jezikovni par natančnost nekoliko nižja, kar je morda moč pojasniti z dejstvom, da se številne večbesedne enote v slovenščini na tem področju prevajajo z enobesedno enoto v angleščini, takih primerov pa naš sistem ne zmore zadovoljivo obdelovati. Po drugi strani so bile ustreznice pri vojaškem korpusu izjemno natančne, kar lahko morda pripišemo stabilnosti vojaške terminologije z razmeroma majhno variabilnostjo. Primeri izluščenih enot iz vseh treh podkorpusov so v Tabeli 3.

	Ustreznost	Angleško	Slovensko
Turizem	0.66	active holidays	aktivne počitnice
	0.66	annual occupancy	letna zasedenost
	0.66	historical heritage	zgodovinska dediščina
	0.58	central reservation system	centralni rezervacijski system
	0.57	sustainable development	trajnostni razvoj
	0.50	improvement of recognisability	dvig prepoznavnosti
	0.49	cultural heritage asset	objekt kulturne dediščine
	0.49	average annual growth rate	povprečna letna stopnja rasti
	0.47	tourist destination development	razvoj turističnih destinacij
	0.44	key brand	ključna tržna znamka
Računovodstvo	0.51	cash flow	denarni tok
	0.45	depreciable asset	amortizirljivo sredstvo
	0.42	intangible asset	neopredmeteno sredstvo
	0.41	taxable temporary difference	obdavčljiva začasna razlika
	0.38	adjusted positive difference	preračunana pozitivna razlika
	0.38	disputable receivable	sporna terjatev
	0.37	onerous contract	kočljiva pogodba
	0.36	realizable value	iztržljiva vrednost

	Ustreznost	Angleško	Slovensko
Računovodstvo	0.33	undistributed net profit	nerazdeljeni čisti dobiček
	0.32	hedging instrument	inštrument za varovanje
Vojska	0.66	anti-personnel mine	protipehotna mina
	0.66	cluster bomb	kasetna bomba
	0.64	tank column	tankovska kolona
	0.62	military pension	vojaška pokojnina
	0.62	defense headquarters	štab obrambe
	0.61	peacetime soldier	mirnodobni vojak
	0.58	territorial defense unit	enota teritorialne obrambe
	0.58	infantry weapon	pehotno orožje
	0.55	military diplomatic representative	vojaški diplomatski predstavnik
	0.55	mountain battalion	gorski bataljon

Tabela 3: Primeri izluščenih terminov

4.3 Terminološke variacije in njihova obdelava pri samodejnem luščenju

Čeprav LUIZ nima posebnih skladenjskih in semantičnih pretvorbenih pravil, se pri opisanem postopku obdelave izluščijo tudi terminološke variacije, to pa je pomembna prednost sistema. Metoda vreče ustreznic namreč upošteva vse možne prevode določene besede, obenem pa upošteva razlike v dolžini obeh terminov ter prevodno verjetnost na besedni ravni. Pri spodnjem primeru nam dvojezični leksikon za besedo *odvzem* ponuja ustreznico *confiscation* z verjetnostjo 1.00, za *orožje* pa *weapon* z verjetnostjo 0.91. Skladenjska variacija *confiscation of weapon* se prav tako izlušči, vendar z nižjo stopnjo ustrezanja zaradi razlike v dolžini izvirnega in ciljnega termina.

odvzem orožja	*weapon confiscation*	*0.63*
	confiscation of weapon	*0.47*

Na drugi strani je naslednji primer semantična terminološka variacija, kjer je drugi možni prevod *amortisable amount* nižje ovrednoten, ker je *amortisable* manj verjetna prevodna ustreznica za *amortizirljiv* kot *depreciable*.

amortizirljivi znesek	*depreciable amount*	*0.27*
	amortizable amount	*0.14*

Kadar pa obstajata tako skladenjska kot semantična variacija, se najvišja vrednost ustrezanja pripiše tisti, ki je hkrati najpogostejša in najbolj jedrnata.

denarna postavka	*cash item*	*0.25*
	item of cash	*0.21*
	monetary item	*0.20*

5 UPORABNIKI LUŠČILNIKA TERMINOLOGIJE

Področje samodejnega luščenja terminologije se tradicionalno povezuje z iskanjem podatkov (*Information Retrieval*), to pa se v zadnjem desetletju pospešeno razvija v smeri semantičnih tehnologij in ontologij. Tako ni presenetljivo, da sodobno pojmovanje terminologije v ospredje postavlja njeno vlogo prenosnika znanja v okviru inteligentnih sistemov in tehnologij znanja. Po drugi strani pa še vedno obstaja tudi bolj primarna skupina uporabnikov, ki si lahko od samodejnega luščenja terminologije – še posebej v večjezikovnem kontekstu – obeta dragoceno podporo, in sicer prevajalci in terminografi.

V času od razvoja prve različice leta 2003 smo LUIZ uporabili za številne naloge, denimo kot podporo pri izdelavi večjezičnega terminološkega slovarja vojaških izrazov, pri gradnji specializiranih terminoloških zbirk za prevajalce v slovenskih vladnih službah in organih EU, pri raziskavi terminotvornih procesov na področju odnosov z javnostmi, pri dograjevanju slovenskega wordneta z večbesednimi enotami (Vintar in Fišer 2008) ter pri razširjanju obstoječega spletnega slovarja informatike z novimi izrazi.[4] Od leta 2008 je poskusna različica enojezičnega luščilnika za slovenščino na voljo tudi na spletu, s čimer smo pridobili še širši krog uporabnikov ter povratnih informacij.

V naštetih uporabnih nalogah so sodelovali različni tipi uporabnikov terminologije: pri gradnji specializiranih slovarjev terminografi in strokovnjaki, pri gradnji terminoloških baz prevajalci in prevajalsko-usmerjeni terminografi, pri jezikovnotehnoloških eksperimentih pa jezikoslovci in računalniški jezikoslovci. Čeprav so bile izkušnje z luščenjem terminologije povečini pozitivne, pa so potrebe in specifične zahteve vseh teh skupin uporabnikov različne in jim z našim sistemom ni bilo mogoče vselej v celoti ugoditi. V naslednjih nekaj odstavkih povzemamo pridobljene izkušnje, predvsem kar se tiče pridobivanja virov za luščenje terminologije ter kakovosti in uporabnosti rezultatov.

[4] http://www.islovar.org

5.1 Luščenje in terminografija

LUIZ smo uporabili v fazi izdelave geslovnika in zbiranja gradiva za tiskano izdajo slovarja vojaških izrazov, in sicer za eno- in dvojezično luščenje terminologije iz vzporednih, primerljivih in enojezičnih korpusov vojaških besedil. Uporabnike je v tem primeru sestavljala skupina profesionalnih terminografov, ki so sami zgradili tudi vse korpuse in so bili dejavno vključeni v prilagajanje luščilnika njihovim potrebam. Ker so bili samodejno izluščeni spiski terminoloških kandidatov namenjeni zgolj kot podlaga ročnemu terminografskemu delu in izbiri terminov, so uporabniki denimo želeli ločene spiske za vsak oblikoskladenjski vzorec, ki so jih nato postopoma obdelovali, se pravi najprej le enobesedne termine, nato le besedne zveze tipa pridevnik + samostalnik, nato le kratice in imena itd.

Število terminov, ki so jih slovaropisci na koncu uvrstili v slovar, je bilo seveda bistveno manjše od števila vseh izluščenih terminov, vendar so bili uporabniki z delovanjem sistema izjemno zadovoljni. Namesto zamudnega ročnega brskanja po obsežnih korpusih ter iskanja prevodnih ustreznic so lahko več časa posvetili pojmovni strukturi vojaške stroke ter opisu slovarskih gesel. Poleg tega so izrazili prepričanje, da je s pomočjo samodejnega luščenja mogoče doseči boljše pokrivanje izrazja izbrane stroke v slovarju. Sodeč po tej izkušnji je luščilnik lahko učinkovito podporno orodje za slovaropisce, pri tem pa natančnost sistema - v razumnih mejah – ne igra odločilne vloge. Večjo težo v takšni situaciji ima priklic oziroma sposobnost sistema, da izlušči tudi redkejše strokovne izraze. Če bi namreč terminograf kljub luščilniku še vedno moral ročno pregledovati gradivo in iskati izraze, ki jih je sistem spregledal, postane smiselnost uporabe luščilnika vprašljiva.

5.2 Luščenje in prevajanje

Prevajalci predstavljajo pomembno skupino uporabnikov terminologije, razširjenost prevajalskih namizij pa je povzročila, da so vzporedni korpusi pravzaprav vsakodnevni stranski proizvod prevajalskega dela. Kljub temu je v številnih prevajalskih okoljih upravljanje terminologije prepuščeno že tako preobremenjenim prevajalcem, zato se številni projekti prevajajo brez sistematične terminološke podpore, če izvzamemo uporabo pomnilnika prevodov. Tudi v prevajalskih okoljih, kjer izdelavi in vzdrževanju terminoloških baz posvečajo potrebno pozornost – denimo pri Službi Vlade RS za razvoj in evropske zadeve (SVREZ), kjer vzdržujejo bazo Evroterm –, je časovni pritisk še vedno odločilni dejavnik, zaradi katerega je ukvarjanje s terminologijo pogosto potisnjeno na zadnje prioritetno mesto.

Sistem LUIZ smo v sodelovanju s Svrezom preskusili dvakrat, obakrat naj bi samodejno luščenje poklicnemu terminografu pomagalo pri gradnji prevajalcem namenjene terminološke baze. Obakrat smo korpus zgradili iz pomnilnika prevodov in ga uporabili za dvojezično luščenje. Zaradi zgoraj omenjenih dejavnikov med samim luščenjem ni bilo posebnega sodelovanja z uporabniki, prav tako sistema nismo posebej prilagajali. Čeprav je bil v obeh primerih odziv terminologa na izluščene spiske načeloma pozitiven, sistem ni v celoti izpolnil pričakovanj, saj je obdelava samodejno izluščenih seznamov zahtevala še precej ročnega dela.

Iz teh izkušenj lahko razberemo, da prevajalska okolja sicer nudijo obilico dvojezičnih virov, primernih za luščenje terminologije, vendar je zaradi tesnih rokov in drugih prioritet za upravljanje terminologije tipično na voljo premalo časa in človeških virov. Kakovostni luščilniki so za prevajalce sicer zagotovo dragocena tehnologija, vseeno pa ne morejo povsem nadomestiti nujnega sistematičnega ukvarjanja s terminologijo.

5.3 Luščenje v jezikoslovju in računalniškem jezikoslovju

V prej opisanih eksperimentih luščenje terminologije predstavlja končno tehnologijo, katere rezultati so gradivo za nadaljnjo človeško obdelavo, na področju računalniškega jezikoslovja pa nasprotno predstavlja korak v predobdelavi vhodnih podatkov za druga orodja in algoritme. Sistem LUIZ smo denimo uporabili za nadgrajevanje slovenskega wordneta z večbesednimi enotami (Vintar in Fišer 2008). Kot večjezični korpus smo pri tem uporabili JRC-ACQUIS, metoda vreče ustreznic pa se je izkazala za učinkovito pri iskanju slovenskih ustreznic za večbesedne enote iz angleškega wordneta. Luščilnik LUIZ smo uporabili tudi pri projektu VoiceTran z namenom izboljšave dvojezičnega leksikona za strojno prevajanje govora (Žganec Gros in dr. 2005), vendar natančnost luščenja pri tem eksperimentu ni bila dovolj visoka, da bi bili učinki opazni pri kakovosti prevajalnika. V času pisanja so v teku eksperimenti samodejnega iskanja definicij v besedilih; tu samodejno izluščene termine uporabljamo kot atribute pri strojnem učenju. Podobno kot pri drugih poskusih se tudi tu kaže, da je natančnost luščenja izredno pomembna, če luščilnik uporabljamo v kombinaciji z drugimi jezikovnimi tehnologijami, saj se napake posameznih faz obdelave medsebojno množijo.

Če sklenemo zgornje misli, je luščenje terminologije tehnologija, ki služi različnim tipom končnih uporabnikov, obenem pa predstavlja pomemben korak v številnih jezikovnotehnoloških aplikacijah znanja. Vsaka uporabniška situacija ima svoje posebne zahteve, ki jih je pri snovanju in prilagajanju sistema za luščenje terminologije treba upoštevati, ne le ker je že sama terminološkost izrazito neulo-

vljiv pojem, ampak predvsem ker imata šum in tišina različne učinke v različnih kontekstih uporabe luščilnika.

6 SKLEP

Opisali smo sistem LUIZ, ki iz vzporednih in primerljivih korpusov lušči terminološke kandidate, za dvojezično poravnavo terminoloških enot pa uporablja metodo vreče ustreznic. Med prednostmi opisanega pristopa so učinkovita obravnava terminoloških variacij in prevodnih alternativ, visoka natančnost poravnave ter uporabnost za luščenje izrazja iz nevzporednih korpusov. Predstavili smo tudi niz evalvacijskih eksperimentov na treh strokovnih področjih.

V zadnjem delu članka razpravljamo o uporabniških vidikih sistemov za luščenje terminologije in iz njih izhajajočih dejavnikih, ki vplivajo na zasnovo in prilagajanje teh sistemov. Pri tem sicer ugotavljamo, da generični luščilnik, ki bi bil primeren za vse vrste aplikacij, ne obstaja, po drugi strani pa se izkaže, da je natančnost pomemben dejavnik pri vseh opisanih scenarijih.

V prihodnje nameravamo LUIZ razširjati na druge jezike in ga v perspektivi opremiti še s komponento za odkrivanje definicij v besedilih. Spletna različica, ki je v času pisanja še enojezična, bo prav tako deležna nadaljnjega razvoja in bo predvidoma kmalu ponujala tudi možnost dvojezičnega luščenja, hkrati pa naj bi uporabniku omogočala prilagajanje določenih parametrov.

Viri

Ahmad, K., Davies, A., Fulford, H., in Rogers, M., 1992: What is a term? The semi-automatic extraction of terms from text. Snell-Hornby et al. (ur.): *Translation Studies – an interdiscipline*. Amsterdam/Philadelphia: John Benjamins.

Ahrenberg, L., Andersson, M. in Merkel, M., 1998: A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98) Montreal, August 10-14, 1998*, 29-35.

Ananiadou, S., 1994: A methodology for automatic term recognition. *Proceedings of the 15th Conference on Computational Linguistics - Volume 2* (Kyoto, Japan, August 05 - 09, 1994). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1034-1038.

Bernth, A., McCord, M. in Warburton, K., 2003: Terminology extraction for global content management. *Terminology* 9:1, 71–98.

Bourigault, D., Gonzalez-Mullier, I., Gros, C., 1996: LEXTER, a Natural Language Processing Tool for Terminology Extraction. Gellerstam, M. et al. (ur.): *Euralex '96 Proceedings I-II*. Göteburg: Universität Göteburg, 771-780.

Collier, N., Nobata, C. in Tsujii, J., 2001: Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Terminology* 7:2, 239–257.

Daille, B., 2003: Conceptual structuring through term variations. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18* (Sapporo, Japan). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 9-16.

Erjavec, T., Ignat, C., Pouliquen, B. in Steinberger, R., 2005: Massive multilingual corpus compilation: Acquis Communautaire and totale. *Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland*. 2005, 32-36.

Estopà Bagot, R., 1999: *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Doctoral thesis, Universitat Pompeu Fabra. Barcelona: UPF.

Fung, P. in McKeown, K., 1997: Finding Terminology Translations from Non-parallel Corpora. *5th Annual Workshop on Very Large Corpora*, Hong Kong: Aug 1997, 192-202.

Gaussier, É., 1998: Flow network models for word alignment and terminology extraction from bilingual corpora. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 444-450.

Gaussier, E.,Renders, J.M., Matveeva, I., Goutte, C. in Dejean, H., 2004: A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04),* 526-533.

Heid, U., 1998: A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology* 5:2, 161 ff.

Hiemstra, D., 1998: Multilingual Domain Modelling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. Coppen, Peter-Arno et al., (ur.): *Proceedings of the 8th CLIN meeting,* 41-58.

Jacquemin, C., 2001: *Spotting and Discovering Terms through Natural Language Processing.* Cambridge/Massachussetts: MIT Press.

Kageura, K., Yoshioka, M., Takeuchi, K., Koyama, T., Tsuji, K. in Yoshikane, F., 2000: Recent advances in automatic term recognition: Experiences from the NTCIR workshop on information retrieval and term recognition. *Terminology* 6:2, 151-174.

Kageura, K., Daille, B., Nakagawa, H. in Chien, L.-F., 2004: Introduction: *Recent trends in computational terminology.* Recent Trends in Computational Terminology, Kageura, K., Daille, B., Nakagawa, H. in Chien, L.-F. (eds.). Amsterdam: John Benjamins, 1–21.

Kwong, Oi Yee, Benjamin K. Tsou in Tom B. Y. Lai, 2004: Alignment and extraction of bilingual legal terminology from context profiles. *Recent Trends in Computational Terminology*, Kageura, K., Daille, B., Nakagawa, H. in Chien, L.-F. (ur.). Amsterdam: John Benjamins, 81–99.

Mann, G. S. in Yarowsky, D., 2001: Multipath translation lexicon induction via bridge languages. *Second Meeting of the North American Chapter of the Association For Computational Linguistics on Language Technologies 2001* (Pittsburgh, Pennsylvania, June 01 - 07, 2001). North American Chapter Of The Association For Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-8.

Melamed, D., 2000: Models of Translation Equivalence between Words. *Computational Linguistics 26(2),* 221-249.

Mima, H. in Ananiadou, S., 2000: An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *Terminology* 6:2, 175–194.

Mima, H., Ananiadou, S., in Matsushima, K., 2006: Terminology-based knowledge mining for new knowledge discovery. *ACM Transactions on Asian Language Information Processing (TALIP)* 5, 1 (Mar. 2006), 74-88.

Nakagawa, H., 2000: Automatic term recognition based on statistics of compound nouns. *Terminology* 6:2, 195–210.

Nenadić, G., Spasić, I., in Ananiadou, S., 2002: Automatic discovery of term similarities using pattern mining. *Coling-02 on COMPUTERM 2002: Second international Workshop on Computational Terminology - Volume 14* International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-7.

Oh, J.-H., Lee, J., Lee, K.-S. in Choi, K.-S., 2000: Japanese term extraction using dictionary hierarchy and machine translation system. *Terminology* 6:2, 287–311.

Tiedemann, J., 2001: Can bilingual word alignment improve monolingual phrasal term extraction? *Terminology* 7:2, 199–215.

Tiedemann, J., 2003: *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Doctoral Thesis, Uppsala: Studia Linguistica Upsaliensia 1.

Uchimoto, K., Sekine, S., Murata, M., Ozaku H. in Isahara, H., 2000: Term recognition using corpora from different fields. *Terminology* 6:2, 233–256.

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

Vintar, Š., 2004: Comparative Evaluation of C-value in the Treatment of Nested Terms. *Memura 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications* (LREC 2004), 54-57.

Vintar, Š. in Fišer, D., 2008: Harvesting multi-word expressions from parallel corpora. *Proceedings of the Language Resources and Evaluation Conference* (LREC 2008), Marrakech, Morrocco, ELRA/ELDA.

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

Vivaldi, J. in Rodríguez, H., 2007: Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13:2, 225–248.

Žganec Gros, J., Mihelič, F., Erjavec, T. in Vintar, Š., 2005: The VoiceTran speech-to-speech communicator. *Text, Speech and Dialogue 2005 (Lecture Notes in Computer Science)*, Berlin: Springer, 379-384.