*Simon KREK**

TERMINOLOGIJA KOT STORITEV**

Povzetek. *V prispevku predlagamo oblikovanje osrednjega slovenskega terminološkega portala, pri katerem bi bili uporabljeni univerzitetni repozitorji oz. digitalne knjižnice, ki vsebujejo strokovna besedila, od diplom, magisterijev in doktoratov do znanstvenih monografij, revij, univerzitetnih učbenikov in druga relevantna besedila. S spletnim sistemom terminologije kot storitve bi bilo mogoče izkoristiti nove računalniške in jezikovne tehnologije, razvite za slovenski jezik, ter zadovoljiti potrebe različnih uporabnikov po takojšnjem dostopu do terminoloških podatkov ter sodelovanju različnih znanstvenih, strokovnih in drugih skupnosti pri gradnji terminoloških virov.*

Ključni pojmi: *terminologija, terminološki portal, procesiranje naravnega jezika, luščenje jezikovnih podatkov, specializirani korpusi, terminološki slovarji*

Uvod

V prispevku se lotevamo vprašanja, kako načrtovati upravljanje s terminologijo na ravni celotne jezikovne skupnosti, v našem primeru slovenske, z uporabo novih računalniških oz. jezikovnih tehnologij, med katerimi so tehnologije luščenja terminoloških kandidatov in drugih relevantnih jezikovnih podatkov iz specializiranih korpusov ter množičenje (*crowdsourcing*) kot nova možnost soudeležbe pri sestavljanju terminoloških virov. Pri predlaganem osrednjem terminološkem portalu je predvidena tudi uporaba pred kratkim vzpostavljenih univerzitetnih ter drugih repozitorijev oz. digitalnih knjižnic, ki vsebujejo strokovna besedila: od diplom, magisterijev in doktoratov do znanstvenih monografij, revij, univerzitetnih učbenikov in drugih relevantnih besedil (prim. Brezovnik in Ojsteršek, 2011). S sistemom terminologije kot storitve (*Terminology as a Service*) bi bilo mogoče izkoristiti nove tehnologije in zadovoljiti potrebe različnih uporabnikov po takojšnjem dostopu do terminoloških podatkov in sodelovanju različnih znanstvenih, strokovnih in drugih skupnosti pri gradnji terminoloških virov.

* *Dr. Simon Krek, Institut »Jožef Stefan«, Ljubljana.*

** *Pregledni znanstveni članek.*

Ta način upravljanja s strokovnim izrazjem je primeren za jezike z manjšim številom govorcev, kar izhaja iz prednosti in slabosti njihove jezikovne situacije. Slabost je predvsem v tem, da je pri teh jezikih finančnih in človeških virov manj, četudi ne glede na število govorcev vsak jezik potrebuje enak terminološki aparat. Prednost pa je (lahko) v tem, da je zaradi majhnosti načeloma lažje organizirati centralizirano storitev za celotno jezikovno skupnost. Kot okvir tovrstnih razmišljanj in prizadevanj nam v nadaljevanju služijo Latvijski terminološki portal *TermNet.lv*,[1] finski projekt *TEPA-termipankkiin*,[2] Slovaška terminološka baza podatkov,[3] hrvaški portal Struna (*Hrvatsko strukovno nazivlje*)[4] in sorodni portali.[5]

Terminološki portali pri drugih jezikih

Vzpostavljanje nacionalnih terminoloških spletnih portalov je bilo značilno že za prvo verzijo spleta (Web 1.0), torej množico spletnih strani z enostranskim komuniciranjem brez možnosti udeležbe uporabnikov spleta pri soustvarjanju vsebine spletnih strani. Portalske spletne rešitve v povezavi s centralnim upravljanjem terminologij pri manjših (evropskih) jezikih se tipično začenjajo vzpostavljati nekje na sredini prejšnjega desetletja, tj. okoli leta 2005. Takrat se je denimo začel eden večjih baltskih terminoloških projektov, ki ga je financirala Evropska komisija v okviru mehanizma eContents, EuroTermBank (*Collection of Pan-European Terminology Resources through Cooperation of Terminology Institutions*). Ta projekt je v marsičem vplival na terminološke portale treh baltskih jezikov. Na spletni strani Slovaške terminološke baze se je mogoče poučiti, da se je slovaški projekt začel jeseni 2005, na strani hrvaškega terminološkega portala Struna, da so se prizadevanja začela leta 2006. Finski projekt TEPA je financiralo finsko Ministrstvo za izobraževanje in kulturo v letih 2006–2007.

Za omenjene projekte oz. portale je značilno, da je v tistem času šlo predvsem za prizadevanje po harmonizaciji terminoloških procesov (leta 2004 je v EU vstopalo večje število držav z nacionalnimi jeziki z manjšim številom govorcev), za zagotavljanje možnosti izmenjave v taki ali drugačni obliki že obstoječih terminoloških podatkov in virov, navadno pa tudi za vzpostavljanje mreže s terminologijo povezanih ustanov in organizacij. Cilj projekta EuroTermBank je bil tako »razvoj in implementacija spletne terminološke banke podatkov, ki naj zagotovi preprost dostop do centraliziranih

[1] *Dostopno preko http://termnet.lv/, 23. 5. 2014.*

[2] *Dostopno preko http://www.tsk.fi/tepa/netmot.exe?UI=engr&height=161, 23. 5. 2014.*

[3] *Dostopno preko http://data.juls.savba.sk/std/ProjectInformation, 23. 5. 2014.*

[4] *Dostopno preko http://struna.ihjj.hr/#, 23. 5. 2014.*

[5] *Hiter pregled se nahaja na hrvaškem terminološkem portalu Struna; dostopno preko http://struna.ihjj.hr/ page/poveznice/, 23. 5. 2014.*

terminoloških virov« (Rirdance in Vasiļjevs, 2006: 11) oz. »konsolidacija terminoloških vsebin, ki izhajajo iz različnih virov in pripadajo različnim lastnikom, z namenom ustvarjanja nacionalnih terminoloških baz podatkov /.../« (*ibid.*). Vse omenjeno se je tipično dogajalo v uveljavljenih institucionalnih okvirih, udeležba širših spletnih skupnosti, bodisi terminoloških ali splošnih, je bila pri teh projektih oz. portalih upoštevana, vendar navadno še nerazvita in močno restriktivna.

V vmesnem času se je vzpostavila in razširila t. i. druga verzija spleta (Web 2.0), za katerega je značilna predvsem intenzivna soudeležba uporabnikov spleta pri ustvarjanju vsebin. Najbolj značilna uspešna zgodba tega razvoja je (bila) Wikipedija, ki je tudi precej popularizirala tip programske opreme wiki, ki omogoča skupinsko ustvarjanje spletnih portalov s pomočjo splošno razširjenih spletnih brskalnikov. V okviru upravljanja s terminologijami se ta razvoj izkazuje v letih pred oz. okoli 2010, ko se je denimo začel projekt Finske terminološke banke znanosti in umetnosti *Tieteen kansallinen termipankki.* Za razliko od predhodnega že omenjenega finskega projekta gre v primeru novega projekta za sistem wiki, okoli katerega se organizira in ga upravlja širša (spletna) terminološka skupnost. Projekt se samoopredeljuje na naslednji način:

> *Finska terminološka banka umetnosti in znanosti je multidisciplinarni projekt, katerega cilj je oblikovanje stalne terminološke baze podatkov za vsa raziskovalna področja na Finskem. V okviru projekta je bila oblikovana platforma Semantic MediaWiki, ki ponuja sodelovalno okolje. To pomeni, da jo vsak lahko prosto uporablja in sodeluje pri diskusijah o terminih.*[6]

Gre za klasični spletni wiki, projekt pa združuje pet pomembnih finskih ustanov, ki se tako ali drugače ukvarjajo z raziskovanjem, terminologijo, finščino in drugim jeziki na Finskem (*Academy of Finland, University of Helsinki, Federation of Finnish Learned Societies, Research Institute of Languages in Finland, The Finnish Terminology Centre*). To kaže, da je sistem wiki postal sprejemljiv tudi kot platforma za akademske skupnosti, skupaj s sistemom množičenja, ki ga predpostavljajo ti sistemi.

Podobno kot finski projekt tudi Slovaška terminološka baza podatkov uporablja sistem wiki, in sicer varianto z imenom MoinMoin.[7] Za razliko od finskega portala, ki v času vzpostavljanja in preizkušanja že vsebuje 24.889 terminoloških vnosov in izkazuje razmeroma intenzivno dejavnost, se zdi, da s trenutnimi približno 5.000 termini slovaški portal v resnici ni zaživel,

[6] *Dostopno preko http://tieteentermipankki.fi/wiki/Termipankki:Etusivu/en, 23. 5. 2014.*

[7] *Dostopno preko http://data.juls.savba.sk/std/ProjectInformation, 23. 5. 2014.*

saj je bilo podobno število omenjeno že pred leti na začetku projekta. O razlogih lahko zgolj ugibamo, zdi pa se, da je pri projektih, ki vključujejo (terminološko) množičenje, pomembno, da je vanje vključeno večje število ustanov, predvsem univerzitetnih, ki lahko pridobijo oz. aktivirajo več uporabnikov, ki potem prispevajo v skupni portal.

V tem delu prispevka smo opisali nekaj značilnih primerov terminoloških portalov pri manjših evropskih jezikih. Tendence pri njihovi organiziranosti kažejo, da portali, ki so bili zasnovani v sredini prejšnjega desetletja, tipično skušajo ponujati čim bolj vseobsegajočo zbirko terminologij različnih področij na enem mestu, portale pa navadno upravljajo nacionalne ustanove oz. konzorciji nacionalnih ustanov. Novejši trend je množičenje, pri katerem s pomočjo sistemov wiki akademska združenja skušajo pritegniti v upravljanje terminologij tudi širšo (znanstveno) skupnost.

Terminološki portali v Sloveniji

Tudi Slovenije omenjeni razvoj ni obšel. Kot pri drugih državah in jezikih, ki so leta 2004 vstopali v Evropsko unijo, je prevajanje zakonodaje EU predstavljalo močan impulz za vzpostavljanje prvega terminološkega spletnega portala z imenom Evroterm, ki je nastal iz zbirke omenjenih prevodov, vseboval pa je tudi korpusni del – Evrokorpus – v spletnem konkordančniku (Željko, 2004; 2009).[8] Prva verzija portala je bila postavljena že leta 2000 in je ves čas vzdrževana in nadgrajevana. Kot je navedeno na strani portala, je trenutno v terminološki zbirki 125.807 vpisov, v korpusih pa več kakor 240 milijonov besed.[9] Poleg evropske zakonodaje so bili v zbirko sčasoma vključeni tudi drugi terminološki viri, ob tem pa je portal tesno povezan tudi z osrednjo terminološko bazo podatkov organov Evropske unije IATE (*InterActive Terminology for Europe*).[10] To pomeni, da je Evroterm trenutno eden od osrednjih spletnih terminoloških virov za slovenščino.

Drugi prosto dostopni spletni terminološki portal z imenom Termania[11] (Krek, 2010; Romih in Krek, 2012) je izdelalo podjetje Amebis, posredno pa ima izvor tudi v raziskovalnem projektu Slovenski terminološki portal (Gorjanc, 2009), ki so ga v letih 2007–2009 izvajali Filozofska fakulteta in Fakulteta za družbene vede Univerze v Ljubljani ter Institut »Jožef Stefan«, sofinanciralo pa ga je že omenjeno podjetje. Na portalu Termania, ki sicer vsebuje vse vrste slovarjev, enojezične, dvojezične, terminološke itd., je aprila 2014 skupaj 43 slovarjev, od tega 18 terminoloških. Pri portalu Termania je treba omeniti, da je za razliko od Evroterma, ki je v celoti zaprt za urejanje vsebin,

[8] *Dostopno preko http://www.evroterm.gov.si/, 23. 5. 2014.*

[9] *Dostopno preko www.evroterm.gov.si/zgodovina.htm, 23. 5. 2014.*

[10] *Dostopno preko http://iate.europa.eu/, 23.5. 2014.*

[11] *Dostopno preko http://www.termania.net/, 23. 5. 2014.*

vanj integriran tudi spletni slovarski urejevalnik, kar pomeni, da ponuja sodelovalno okolje, ki omogoča tudi množičenje, čeprav ne v smislu klasičnega sistema wiki. Če bi želeli vzpostaviti sodelovalno spletno okolje, ta lastnost pomeni precejšnjo dodano vrednost pri izbiri.

Tretje spletno mesto, ki ga le pogojno lahko imenujemo terminološki portal v pomenu, kot ga uporabljamo v prispevku, je Terminologišče Inštituta za slovenski jezik Frana Ramovša na ZRC SAZU.[12] Stran med drugim vsebuje spletne iskalnike po izboru v knjižni obliki objavljenih slovarjev terminološke sekcije Inštituta. Vsak spletno dostopni slovar ima svoj, samostojen iskalnik, kar je sicer v nasprotju z načeli spletnih terminoloških bank, ki težijo k poenostavljanju dostopa do podatkov iz čim večjega števila terminoloških baz. Z izbiro takega načina dostopa do podatkov je poudarjena med drugim tudi zaključenost in nespremenljivost vsebin, kar pomeni, da se organizacija portala izrazito prilagaja logiki tiskanih, knjižnih izdaj.

Po drugi strani Terminologišče poleg iskanja po slovarjih ponuja tudi storitev, ki je drugi portali nimajo – terminološko svetovanje. V okviru svetovalnice je na portalu trenutno objavljenih nekaj manj kot 50 odgovorov, ki časovno segajo od leta 2011 do 2014. Osnovni način komuniciranja, ki se uporablja v okviru svetovalnice, je elektronska pošta, kar poleg ostalih odločitev kaže, da spletna stran nima ambicije, da bi delovala kot vozlišče za izmenjavo informacij med različnimi terminološkimi skupnostmi v Sloveniji, temveč gre bolj za izolirano spletno mesto Sekcije za terminološke slovarje Inštituta za slovenski jezik Frana Ramovša ZRC SAZU.

Poleg omenjenih portalov na slovenskem spletu obstaja cela vrsta samostojnih terminoloških slovarjev, zbirk in podobnih virov. Večina relevantnih povezav je zbrana na strani Spletni slovarji,[13] ki se nahaja v okviru portala Evroterm in je redno posodabljana od leta 2001. Med samostojnimi terminološkimi spletnimi mesti je treba izpostaviti predvsem Islovar (Puc, 2009; Turk in Puc, 2007),[14] ki ga izdaja jezikovna sekcija Slovenskega društva Informatika in vsebuje 6.550 izrazov s področij informatike, informacijske tehnologije in telekomunikacij. Tako kot Termania tudi ta portal omogoča urejanje slovarja v spletnih brskalnikih, poleg tega lahko nove izraze vnašajo vsi registrirani uporabniki, kar pomeni, da omogoča široko uporabniško soudeležbo pri urejanju. Pri tem sicer ne gre za klasični sistem množičenja, saj uporabniki nove izraze lahko le predlagajo, potem pa ti vstopijo v nadzorovan uredniški proces. Ta predpostavlja, da se slovarsko geslo pomika po poti od predloga do končnega potrjenega izraza oz. prevoda, predvidena pa je uporaba uredniških značk: predlog, pregledano, strokovno pregledano in

[12] *Dostopno preko http://isjfr.zrc-sazu.si/terminologisce#v, 23. 5. 2014.*

[13] *Dostopno preko http://www.evroterm.gov.si/slovar/index.html, 23. 5. 2014.*

[14] *Dostopno preko http://www.islovar.org/, 23. 5. 2014.*

urejeno. Take in podobne sisteme urejanja poznajo tudi drugi, tujejezični terminološki spletni portali in predstavljajo uveljavljeno prakso.

Zaključimo lahko z ugotovitvijo, da v Sloveniji dejavnosti, povezane z organizacijo, dostopom in upravljanjem terminologij v digitalnem oz. spletnem okolju – tako kot pri drugih jezikih s podobnim številom govorcev – potekajo že dlje časa. Če privzamemo, da je cilj participacija čim širše (znanstvene) skupnosti in da tehnološko sodelovalna spletna okolja niso več problem, lahko nadalje ugotovimo, da prizadevanja po vzpostavitvi osrednjega spletnega mesta, kjer bi posamezna znanstvena veda ali terminološka skupina lahko samoorganizirala svojo terminološko dejavnost, do sedaj niso bila uspešna.

Obstajajo trije ali štirje kandidati za nadgrajevanje v bodoči vseslovenski terminološki portal, vendar ima vsak svoje slabosti. Pri Evrotermu gre za zaprt sistem, pri katerem je vzpostavitev spletnega sodelovalnega okolja za širšo skupnost verjetno težko izvedljiva. Pri Termanii je zadrega predvsem dolgoročno omogočanje proste dostopnosti storitve, saj je vzdrževanje sistema v celoti odvisno od altruističnega vlaganja zasebnega podjetja v po svoji naravi javni servis. Terminologišče kot skupek iskalnikov po posameznih slovarjih ne izkazuje ambicije po vzpostavljanju centralne terminološke banke, v okviru centralne storitve pa bi bila vsekakor koristna njegova terminološka svetovalnica. Pri Islovarju kot dodatnem kandidatu je vprašanje predvsem možnost obvladovanja bistveno večje količine slovarjev in njihovih podatkov ter uporabniške soudeležbe.

Leta 2010 eden od deležnikov pri dosedanjih tovrstnih prizadevanjih piše: »V slovenskem prostoru je bilo v zadnjem času narejenih kar nekaj korakov k tesnejšemu povezovanju različnih akterjev na področju terminološkega dela, v prihodnje pa bi bilo treba nadaljevati v smeri iskanja čim večjih skupnih sinergijskih učinkov« (Gorjanc, 2010: 102). Zdi se, da v vmesnem času do preboja sinergije še ni prišlo, ključno vprašanje pa je, ali je mogoče iz omenjenih portalov dejansko sestaviti vseslovenski terminološki portal kot centralno točko za upravljanje s terminologijami za vse znanstvene, strokovne in druge skupnosti, ki te podatke potrebujejo.

Nacionalni portal odprte znanosti

Nedavni razvoj prosto dostopnih univerzitetnih repozitorijev in digitalnih knjižnic v Sloveniji daje ideji osrednjega slovenskega terminološkega portala dodatno dimenzijo. V letu 2013 je bil dokončan projekt vzpostavitve Nacionalnega portala odprte znanosti,[15] ki vključuje Digitalno knjižnico

[15] *Dostopno preko http://openscience.si/, 23. 5. 2014.*

Univerze v Mariboru (KDUM),[16] Repozitorij Univerze v Ljubljani (RUL),[17] Repozitorij Univerze na Primorskem (RUP)[18] in Repozitorij Univerze v Novi Gorici.[19] Po navedbah na skupnem portalu je v repozitorijih nekaj več kot 61.000 dokumentov, med njimi predvsem diplome, magisteriji in doktorati, tudi znanstvene revije, skripta ter druga znanstvena in strokovna dela.

Prizadevanje glede vzpostavitve univerzitetnih rezpozitorijev je sicer močno povezano z željami širše raziskovalne skupnosti ter tudi državnih in teles EU, ki raziskave financirajo, da bi bili rezultati javno financiranih raziskovalnih dejavnosti prosto dostopni na spletu.[20] Nekje od 18. oz. 19. stoletja do širitve svetovnega spleta je namreč izmenjava znanstvenih izsledkov potekala prek objav v tiskanih znanstvenih revijah, ki so bile vpete v klasični založniški krog z naročninami, kar pomeni, da so bili raziskovalni rezultati plačani vsaj dvakrat: s financiranjem raziskav in naročninami. Na svetovnem spletu pa je tovrstne podatke mogoče deliti pod pogoji prostega dostopa, če so primerno urejene avtorske pravice, kar je spodbudilo razmišljanja o 'odprti znanosti'.

V Sloveniji velik del znanstvenih revij sofinancira Javna agencija za knjigo, ki je pred časom naročila ekspertizo z naslovom Izvedbeni načrt spletnega portala znanstvenih in literarnih revij (Breznik et al., 2011), ki vsebuje »vsebinski, organizacijski, finančni in izvedbeni načrt za zbirni portal 148 znanstvenih in literarnih revij, ki jih ta čas financira Javna agencija za knjigo RS«. Nadalje, priporočilo Evropske komisije z naslovom *Boljšemu dostopu do znanstvenih informacij naproti: izboljšanje učinkovitosti javnih naložb v raziskave*, sprejeto leta 2012, si postavlja tri cilje, ki kažejo na dolgoročnejši premik večine javno financiranih raziskovalnih rezultatov v okvir odprtega dostopa:

- Do leta 2014 bodo v vseh državah članicah in na vseh ustreznih ravneh vzpostavljene politike za odprti dostop do znanstvenih člankov in podatkov.
- Delež iz javnih sredstev financiranih znanstvenih člankov, ki bodo prek odprtega dostopa na voljo po vsej EU, se bo do leta 2016 povečal z 20 % na 60 %.
- 100 % znanstvenih objav, nastalih v okviru programa Obzorje 2020, bo na voljo prek odprtega dostopa. (Evropska komisija, 2012: 12)

V kontekstu naslovne teme prispevka sta odprti dostop in prizadevanja po vzpostavitvi univerzitetnih repozitorijev in portalov znanstvenih revij

[16] *Dostopno preko http://dkum.uni-mb.si/, 23. 5. 2014.*

[17] *Dostopno preko http://repozitorij.uni-lj.si/, 23. 5. 2014.*

[18] *Dostopno preko http://www.odun.univerza.si/info/index.php/slo/, 23. 5. 2014.*

[19] *Dostopno preko http://repozitorij.ung.si/, 23. 5. 2014.*

[20] *Dostopno preko http://www.openaccess.si/, 23. 5. 2014.*

pomembna zato, ker ponujata možnost za vzpostavitev osrednjega terminološkega portala, v okviru katerega bi bilo mogoče iz zbranih znanstvenih in strokovnih del izdelati specializirane korpuse po področjih, vedah ali poljubnih drugih kriterijih in uporabiti novo razvite računalniške in jezikovne tehnologije, ki bi bistveno olajšale tradicionalno terminološko delo.

Računalniške in jezikovne tehnologije

Za slovenščino so bile v zadnjih letih razvite tudi računalniške in jezikovne tehnologije, ki jih je smiselno uporabiti ob vzpostavitvi terminološkega portala. Prvi pogoj za uporabo teh tehnologij je pretvorba obstoječih besedil v množico specializiranih besedilnih korpusov (prim. Arhar Holdt, 2006), kakršne poznamo iz različnih preteklih projektov, denimo korpus DSI (Erjavec in Vintar, 2004), s pomočjo katerega nastaja Islovar, Korpus vojaških besedil Grizold (Gorjanc in Logar Berginc, 2007) kot podlaga za Vojaški slovar (Pečovnik, 2009), Korpus besedil odnosov z javnostmi KoRP (Kalin Golob in Logar, 2008) kot podlaga za terminološko podatkovno zbirko odnosov z javnostmi TERMIS (Logar, 2013), Večjezični korpus turističnih besedil TURK (Mikolič et al., 2009) kot podlaga za Turistični terminološki slovar (Mikolič et al., 2011) in drugi.

Klasični postopki sestavljanja specializiranega korpusa, od zbiranja besedil do osnovne računalniške obdelave – segmentacija, tokenizacija besedil, lematizacija in oblikoskladenjsko označevanje, dodajanje metapodatkov itd. – so natančneje opisani v omenjenih virih. Če predpostavimo, da bi bilo pri gradnji specializiranih korpusov iz repozitorijev treba obdelati desetine tisočev dokumentov, določen izziv predstavlja predvsem avtomatiziranje postopka, ker je težko predpostaviti kakršenkoli ročni poseg pri obdelavi in pretvorbi besedil iz izvirnih formatov in pri pridobivanju metapodatkov, ki so v repozitorijih sicer večinoma že na voljo v relacijskih podatkovnih bazah. Drugi, morda celo večji izziv je konsistentno avtomatsko razporejanje besedil po smiselnih enotah, denimo znanstvenih vedah ali ožjih področjih. Pri tem so do neke mere sicer lahko v pomoč metapodatki, kot je navedba fakultete, mentorjev, naslovi, ključne besede itd., kljub temu pa bi bilo smiselno v avtomatizirani proces vključiti tudi tehnologije, kot je primerjava podobnosti (Fortuna et al., 2005; Juršič et al., 2013) oz. klasifikacija dokumentov (Mladenić et al. 2011; Brank et al., 2008; Grobelnik in Mladenić, 2005).

Po avtomatskem oblikovanju specializiranih korpusov sledi luščenje terminoloških kandidatov (Vintar, 2009; Logar et al., 2012; Arhar Holdt, 2011) iz posameznega korpusa, kar je pri slovenščini prav tako že razmeroma uveljavljena tehnologija. V primeru enojezičnih specializiranih korpusov navadno gre za primerjavo besedišča, kot ga najdemo v splošnih korpusih,

z besediščem v specializiranem korpusu. Pri tem so v uporabi različne statistične metode, včasih tudi z upoštevanjem tipičnih leksikalno-skladenjskih vzorcev. Za slovenščino obstaja luščilnik terminologije LUIZ (Vintar, 2010) in je bil uporabljen v več projektih, demo je mogoče tudi preizkusiti na spletni strani projekta Slovenski terminološki portal.[21]

Izluščeni terminološki kandidati predstavljajo izhodišče za strojno luščenje drugih leksikalnih podatkov oz. delov terminoloških gesel, ki so pripravljeni za ročno obdelavo gesel v spletnem slovarskem urejevalniku. V okviru projekta Sporazumevanje v slovenskem jeziku[22] je bil nedavno razvit postopek za luščenje statistično izstopajočih in značilnih skladenjskih struktur, kolokacij in (dobrih slovarskih) zgledov, v katerih se pojavlja terminološki kandidat (Kosem et al., 2013a; Kosem et al., 2013b), in sicer s pomočjo orodja Sketch Engine (Kilgarriff et al., 2004).[23] Ta postopek je bil kasneje preizkušen tudi na specializiranem korpusu KoRP v okviru projekta TERMIS (Logar, 2013; Logar in Kosem, 2013). Na ta način se slovarsko oz. terminološko delo iz pisanja preobrazi primarno v urejanje strojno izločenih informacij iz korpusov, pri čemer digitalno zasnovani slovarji lahko ponudijo bistveno več informacij kot klasični tiskani slovarji (prim. Krek et al., 2013; Krek, 2014).

Poleg luščenja skladenjskih struktur, kolokacij in zgledov je mogoče iz korpusov izluščiti tudi definicije, kar je sicer tehnično precej zahteven postopek, ki pa je bil že preizkušen tudi na slovenskem jeziku (Fišer et al., 2010; Pollak et al., 2012). V literaturi je tako navedeno, da je

> *iz korpusa mogoče izluščiti veliko število potencialnih definicij, ki so v približno tretjini primerov prave definicije. Algoritem za klasifikacijo, ki smo ga izurili, je na testnih množicah iz Wikipedije deloval s skoraj 83-odstotno točnostjo, medtem ko smo pri klasifikaciji primerov iz korpusa dosegli 71-odstotno točnost* (Fišer et al., 2011: 149).

Ta postopek je torej mogoče uporabiti tudi za avtomatsko pripravo gesel iz specializiranih korpusov za terminološke slovarje po vedah oz. področjih. Vsem predhodno že preizkušenim postopkom pa bi bilo smiselno dodati tudi postopek prepoznavanja in pretvorbe oz. vnosa obstoječih glosarjev, ki so pogosto del diplom, magisterijev ali doktoratov, v terminološki portal.

[21] *Dostopno preko http://lojze.lugos.si/cgitest/extract.cgi, 23. 5. 2014.*

[22] *Dostopno preko http://www.slovenscina.eu/, 23. 5. 2014.*

[23] *Dostopno preko http://www.sketchengine.co.uk/, 23. 5. 2014.*

Sklep

Enotni slovenski terminološki portal bi torej gradil na dveh osnovah: (1) preteklih prizadevanjih različnih ustanov, od državnih organov, ki so se pred desetimi in več leti soočali z velikim izzivom prevajanja celotne evropske zakonodaje (skupaj s sedanjim slovenskim delom prevajalske službe EU), raziskovalcev na univerzah in inštitutih, ki so se do sedaj ukvarjali s prehodom upravljanja s terminologijami v digitalno okolje, podjetij, ki so vlagala v razvoj terminoloških portalov in vseh drugih deležnikov. Poleg izrabe že obstoječe infrastrukture bi bilo (2) smiselno v portal integrirati novo nastale tehnologije, ki jih denimo pred štirimi ali petimi leti še ni bilo, in deležnike na ustanovah, ki se s temi tehnologijami ukvarjajo.

Če povzamemo, bi osrednji slovenski terminološki portal torej moral omogočati naslednje procese in aktivnosti:

- avtomatsko luščenje terminoloških kandidatov iz obstoječih repozitorijev, digitalnih knjižnic in dokumentov, ki jih na portal naložijo uporabniki;
- avtomatsko prepoznavanje prevodnih ustreznic za avtomatsko izluščene termine iz dvojezičnih korpusov, skritih vzporednih spletnih korpusov in drugih dvo- ali večjezičnih (korpusnih, prevodnih ali slovarskih) virov;
- avtomatsko dodajanje leksiko-gramatičnih in semantičnih podatkov izluščenim terminološkim kandidatom;
- možnost ročnega čiščenja in urejanja strojno izluščenih podatkov v spletnem slovarskem urejevalniku;
- uvoz podatkov v terminološki portal in izvoz iz njega v standardiziranih formatih (TBX, tudi TEI, LMF in drugi).[24]

Shematično predlog opisanega portala lahko prikazuje slika 1.

V evropskem okviru podobne iniciative že obstajajo, ena od značilnih je denimo projekt *TaaS: Terminology as a Service*,[25] ki ga financira Evropska komisija v okviru 7. operativnega programa in se zaključuje v letu 2014. Tako kot v primeru projekta EuroTermBank, ki je bil neke vrste znanilec dobe spletnih terminoloških bank podatkov, lahko predpostavimo, da bo razvoj v prihodnje šel v smer, ki jo nakazuje omenjeni projekt. Predpostavljamo lahko, da je predlagana rešitev nakazana že v Resoluciji o nacionalnem programu za jezikovno politiko 2014–2018.[26] V delu, ki govori o opremljenosti jezika glede terminologije, je kot eden od ciljev navedena

[24] *Term Base eXchange (TBX), Dostopno preko http://www.ttt.org/oscarstandards/tbx/tbx_oscar.pdf, 23. 5. 2014; Text Encoding Initiative (TEI), Dostopno preko http://www.tei-c.org/, 23. 5. 2014; Lexical Markup Framework (LMF), Dostopno preko http://www.lexicalmarkupframework.org/, 23. 5. 2014.*

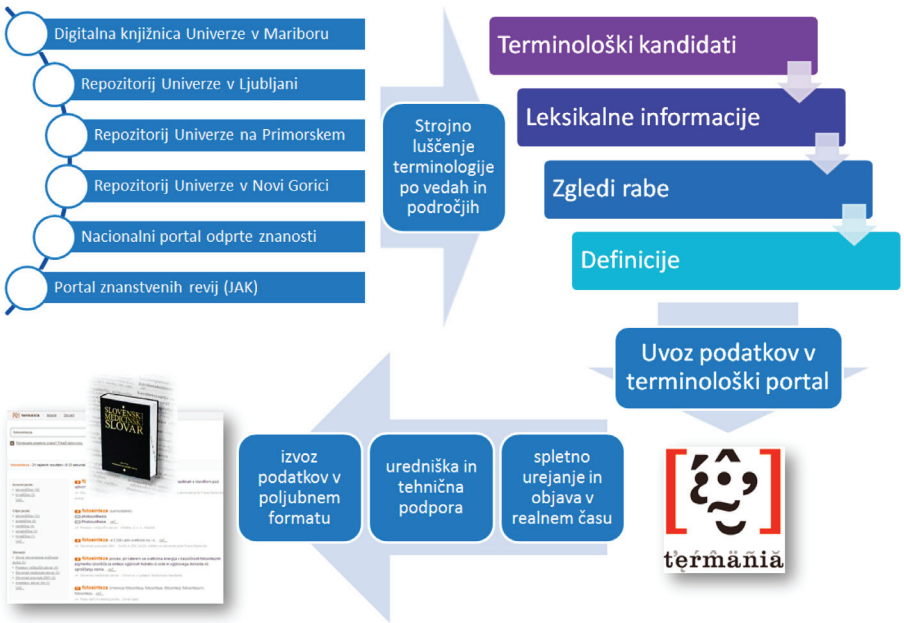
[25] *Dostopno preko http://www.taas-project.eu/, https://demo.taas-project.eu/, 23. 5. 2014.*

[26] *Dostopno preko http://www.pisrs.si/Pis.web/pregledPredpisa?id=RESO91, 23. 5. 2014.*

vzpostavitev terminološkega portala kot dela prosto dostopnega spletnega portala s čim več dosegljivimi jezikoslovnimi podatki o slovenščini, ki bo vseboval tako obstoječe kot tudi nastajajoče terminološke slovarje, terminološke baze in učinkovit svetovalni servis ter bo izkoriščal možnosti hitre izmenjave znanja in mnenj med področnimi strokovnjaki in jezikoslovci, ki jih omogoča splet. V okviru portala mora biti predvidena tudi vzpostavitev nacionalnega mehanizma za potrjevanje terminologije Evropske unije.

Predlagani enotni slovenski terminološki portal želi izpolniti navedeni resolucijski cilj. Pri tem pa je treba upoštevati, da so pri jezikih s podobnim številom govorcev, kot je slovenščina, lahko uspešni le projekti, pri katerih se pri skupnem cilju združi zadostna kritična masa relevantnih ustanov in deležnikov, saj dosedanje stanje jasno kaže, da izolacija ne vodi do optimalnega rezultata. Glede na to, da omenjena resolucija kot nosilce teh aktivnosti izpostavlja Javno agencijo za raziskovanje RS, Ministrstvo za izobraževanje, znanost in šport, Ministrstvo za kulturo, vsi v sodelovanju s SAZU, univerzami in raziskovalnimi inštitucijami, lahko rečemo, da so s tem po vsej verjetnosti identificirane tudi ustanove, ki bi pri projektu morale sodelovati z združenimi močmi, prve kot podporniki, druge pa kot izvajalci.

Slika 1: SHEMA TERMINOLOŠKEGA PORTALA



LITERATURA

Arhar Holdt, Špela (2006): Gradnja specializiranega korpusa. Jezik in slovstvo 51 (1): 53–67 Ljubljana: Slavistično društvo Slovenije.

Arhar Holdt, Špela (2011): Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Brank, Janez, Dunja Mladenić, Marko Grobelnik in Nataša Milić-Frayling (2008): Feature selection for the classification of large document collections. Journal for universal computer science 14 (10): 1562–1596.

Breznik, Maja, Valentin Gjorgjioski in Matija Damjan (2011): Izvedbeni načrt spletnega portala znanstvenih in literarnih revij: ekspertiza. Ljubljana: Javna agencija za knjigo Republike Slovenije.

Brezovnik, Janez in Milan Ojsteršek (2011): Digital library of University of Maribor: (more than just a bunch of documents). V International Conference on Applied Computer Science (ACS), 473–478. Valetta: Institute for Environment, Engineering, Economics and Applied Mathematics.

Erjavec, Tomaž in Špela Vintar (2004): Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika. Uporabna informatika 12 (2): 97–106. Ljubljana: Slovensko društvo Informatika.

Fišer, Darja, Senja Pollak in Špela Vintar (2010): Learning to mine definitions from Slovene structured and unstructured knowledge-rich resources. V Language Resources and Evaluation Conference 2010: 2932–2936. Valletta: Language Resources and Evaluation Conference.

Fišer, Darja, Senja Pollak in Špela Vintar (2011): Samodejno luščenje definicij iz specializiranih besedil. V Simona Kranjc (ur.), Meddisciplinarnost v slovenistiki, 145–150. Ljubljana: Znanstvena založba Filozofske fakultete.

Fortuna, Blaž, Dunja Mladenić in Marko Grobelnik (2005): Visualization of text document corpus. Informatica 29 (4): 497–502. Ljubljana: Slovensko društvo informatika.

Gorjanc, Vojko (2009): Slovenski terminološki portal. V Nina Ledinek (ur.), Mojca Žagar Karer (ur.), Marjeta Humar (ur.), Terminologija in sodobna terminografija, 303–310. Ljubljana: ZRC SAZU.

Gorjanc, Vojko (2010): Terminološko načrtovanje in upravljanje terminologije. Slavistična revija 58 (1): 95–104. Ljubljana: Slavistično društvo Slovenije.

Gorjanc, Vojko in Nataša Logar Berginc (2007): Od splošnih do specializiranih korpusov – načela gradnje glede na njihov namen. V Irena Orel (ur.), Razvoj slovenskega strokovnega jezika, 637–650. Ljubljana: Filozofska fakulteta.

Grobelnik, Marko in Dunja Mladenić (2005): Simple Classification Into Large Topic Ontology of Web Documents. Journal of Computing and Information Technology 13 (4): 279–285. Zagreb: University Computing Centre.

Juršič, Matjaž, Bojan Cestnik, Tanja Urbančič in Nada Lavrač (2013): HCI empowered literature mining for cross-domain knowledge discovery. V Andreas Holzinger (ur.), Gabriella Pasi (ur.), Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Third International Workshop, HCI-KDD 2013, 124–135. Berlin, Heidelberg: Springer.

Kalin Golob in Nataša Monika Logar (2008): Terminologija odnosov z javnostmi: od upoštevanja terminoloških načel do pridobivanja podatkov iz besedil. Teorija in praksa 45 (6): 663–677.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz in David Tugwell (2004): The Sketch Engine. V Geoffrey Williams (ur.), Sandra Vessier (ur.), Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 105–116. Lorient: Université de Bretagne-Sud.

Kosem, Iztok, Polona Gantar in Simon Krek (2013a): Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. V Iztok Kosem (ur.), Jelena Kallas (ur.), Polona Gantar (ur.), Simon Krek (ur.), Margit Langemets (ur.), Maria Tuulik (ur.), Electronic lexicography in the 21st century: proceedings of eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia, 32–48. Ljubljana: Trojina, Zavod za uporabno slovenistiko; Talin: Eesti Keele Instituut.

Kosem, Iztok, Polona Gantar in Simon Krek (2013b): Avtomatizacija leksikografskih postopkov. Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave 1 (2): 139–164. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Krek, Simon (2010): Termania free on-line dictionary portal. V Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010), 928–930. Ljouwert: Fryske Akademy.

Logar, Nataša (2013): Korpusna terminografija: primer odnosov z javnostmi. Ljubljana: Trojina, zavod za uporabno slovenistiko, in Fakulteta za družbene vede.

Logar, Nataša in Iztok Kosem (2013): TERMIS: a corpus-driven approach to compiling an e-dictionary of terminology. V Iztok Kosem (ur.), Jelena Kallas (ur.), Polona Gantar (ur.), Simon Krek (ur.), Margit Langemets (ur.), Maria Tuulik (ur.), Electronic lexicography in the 21st century: proceedings of eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia, 164-178. Ljubljana: Trojina, Zavod za uporabno slovenistiko; Talin: Eesti Keele Instituut.

Logar, Nataša, Špela Vintar in Špela Arhar Holdt (2012): Luščenje terminoloških kandidatov za slovar odnosov z javnostmi. V Tomaž Erjavec (ur.), Jerneja Žganec Gros (ur.), Zbornik Osme konference Jezikovne tehnologije, 8. – 12. oktober 2012, 135–140. Ljubljana: Institut Jožef Stefan.

Mikolič, Vesna, Maja Smotlak, Klara Šumenjak, Jana Volk, Mojca Kompara, Martina Rodela, Elena Šverko in Jernej Vičič (2011): Turistični terminološki slovar. Kamnik: Amebis. Dostopno preko http://www.termania.net/slovarji/78/turisticni-terminoloski-slovar, 2. 5. 2014.

Mikolič, Vesna, Jernej Vičič in Jana Volk (2009): Namen in metode urejanja večjezičnega korpusa turističnih besedil (TURK). V Vesna Mikolič (ur.), Jezikovni korpusi v medkulturni komunikaciji, 65–74. Koper: Univerza na Primorskem.

Mladenic, Dunja, Janez Brank in Marko Grobelnik (2011): Document Classification. V Encyclopedia of Machine Learning, 289–293. New York: Springer.

Pečovnik, Tina (2009): Vojaška terminologija. V Nina Ledinek (ur.), Mojca Žagar Karer (ur.), Marjeta Humar (ur.), Terminologija in sodobna terminografija, 215–225. Ljubljana: ZRC SAZU.

Pollak, Senja, Anže Vavpetič, Janez Kranjc, Nada Lavrač in Špela Vintar (2012): NLP workflow for on-line definition extraction from English and Slovene text corpora. V Proceedings of the Conference on Natural Language Processing 2012/11th Conference on Natural Language Processing (KONVENS), September 2012, 53–60. Dunaj: Österreichischen Gesellschaft für Artificial Intelligende.

Puc, Katarina (2009): Urejanje spletnega terminološkega slovarja Islovar. V Nina Ledinek (ur.), Mojca Žagar Karer (ur.), Marjeta Humar (ur.), Terminologija in sodobna terminografija, 335–343. Ljubljana: ZRC SAZU.

Rirdance, Signe (ur.), Vasiljevs, Andrejs (ur.) (2006): Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project. Riga: EurotermBank Consortium.

Romih, Miro in Simon Krek (2012): Termania – prosto dostopni spletni slovarski portal = Termania – freely available online lexical portal. V Tomaž Erjavec (ur.), Jerneja Žganec Gros (ur.), Zbornik Osme konference Jezikovne tehnologije, 8. – 12. oktober 2012, 163–166. Ljubljana: Institut Jožef Stefan.

Turk, Tomaž in Katarina Puc (2007): Eslovar kot model spletnega terminološkega slovarja. V Irena Orel (ur.), Razvoj slovenskega strokovnega jezika, 651–663. Ljubljana: Filozofska fakulteta.

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

Vintar, Špela (2010): The bag-of-equivalents term alignment approach and its evaluation. Terminology (16) 2: 141–158. Amsterdam, Filadelfija: Benjamins.

Željko, Miran (2004): Evroterm in Evrokorpus – terminološki slovar in korpus prevodov = Evroterm and Evrokorpus – a terminology database and a corpus of translations. V Marjeta Humar (ur.), Terminologija v času globalizacije, 139–149. Ljubljana: ZRC SAZU.

Željko, Miran (2009): Povezava večjezične terminološke zbirke z večjezičnim korpusom. V Nina Ledinek (ur.), Mojca Žagar Karer (ur.), Marjeta Humar (ur.), Terminologija in sodobna terminografija, 329–333. Ljubljana: ZRC SAZU.

VIRI

Evropska komisija (2012): Towards better access to scientific information: Boosting the benefits of public investments in research. Dostopno preko http://ec.europa.eu/ research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf, 2. 5. 2014.

Krek, Simon (2014): SSKJ v slovarski bazi. Dostopno preko http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/5-_Simon_Krek_SSKJ_v_slovarski_bazi.pdf, 2. 5. 2014.

Krek, Simon, Iztok Kosem in Polona Gantar (2013): Predlog za izdelavo Slovarja sodobnega slovenskega jezika. Dostopno preko http://trojina.org/slovar-predlog/datoteke/Predlog_ SSSJ_v1.1.pdf, 2. 5. 2014.