Špela Vintar
Univerza v Ljubljani, Filozofska fakulteta
spela.vintar@guest.arnes.si

UDK 811.163.6'42:81'374:004

KAJ IZVIRA IZ JEZIKOVNIH VIROV

Jezikovni viri, kot so korpusi, leksikoni in tezavri, pa tudi označevalniki in druge temeljne tehnologije, niso sami sebi namen, temveč služijo za izdelavo pomembnih jezikovnih priročnikov in uporabniških aplikacij. Prispevek predstavi načine izrabe jezikovnih virov v jezikoslovne in nejezikoslovne namene s poudarkom na uporabniškem vidiku. Med prvimi omenjamo predvsem korpusne metode pri opisovanju jezika v leksikografiji, slovnici in prevodoslovju, nato sledi pregled širše znanih jezikovnih orodij, kot so črkovalniki, slovarji sopomenk ali prevajalniki. V zadnjem času se razvoj vse bolj osredotoča na pomensko usmerjene aplikacije, na primer razvrščanje in povzemanje dokumentov, iskanje podatkov in gradnjo ontologij. V drugem delu prispevka pobliže opišemo dve področji izrabe korpusnih virov, in sicer luščenje terminologije in pridobivanje medicinskih spoznanj s pomočjo rudarjenja besedil.

1 Uvod

Z nekajletno zamudo za »velikimi« jeziki je tudi pri nas korpusno jezikoslovje zaživelo kot dinamično interdisciplinarno področje raziskovanja in ustvarjanja. Z mejnikom, ki ga predstavlja korpus slovenskega jezika FIDA <http://www.fida.net>, z vse daljšim seznamom drugih korpusnih virov za slovenski jezik in nenazadnje z načrtovanimi in tekočimi projekti nadaljnje izgradnje slovenske jezikovnotehnološke infrastrukture je področje nedvomno preseglo kritično točko adolescence. Obenem pa je zdaj tudi čas, da se izraziteje in bolj sistematično posvetimo ciljem, ki bodo s pomočjo korpusov postali dosegljivi in zaradi katerih se je vse skupaj sploh začelo.

Članek se tako posveča izrabi korpusov in predstavlja le ozek vpogled v nepregleden spekter tehnologij, aplikacij in dejavnosti, ki temeljijo na elektronskih zbirkah besedil in vključujejo precej več kot le z jezikoslovjem povezana področja. V prvem delu pregledno predstavimo nekatera tradicionalna in sodobnejša področja izrabe korpusnih virov, v drugem delu pa nekoliko pobliže opišemo dva načina pridobivanja podatkov iz korpusov, in sicer luščenje terminologije (Term Extraction) in rudarjenje podatkov iz besedil (Text Mining).

2 Korpusi v jezikoslovju

Splošna utemeljitev korpusnega pristopa, kot ga povzema na primer Geoffrey Leech (1991), je, da »omogoča raziskovanje jezika s pomočjo primerov jezikovne rabe iz resničnega življenja.« V skladu s tem je poglavitno in najstarejše področje izrabe korpusov povezano z opisovanjem različnih plati določenega jezika, na primer besedišča, slovnice, stilnih značilnosti, komunikacijsko-funkcionalnih zvrsti in drugih. Čeprav so jezikovni opisi od nekdaj črpali iz besedilnih virov in pogosto vključevali podrobno in dolgotrajno pregledovanje gradiva, so računalniški korpusi prinesli metodološki preporod v malone vse veje jezikoslovja. Tako si danes nekaterih, predvsem leksikografskih opravil brez ustrezne računalniške podpore sploh ne moremo več zamišljati, z vse večjimi korpusi in boljšimi načini avtomatskega označevanja pa je mogoče avtomatsko pridobiti korpusne dokaze tudi za kompleksnejše jezikovne pojave, na primer kohezivnost besedil ali stilne posebnosti.

V nadaljevanju naštejemo le nekaj korpusnih metod, ki so korenito spremenile dotlej uveljavljene jezikoslovne prakse (glej tudi Thomas in Short 1996; McEnery in Wilson 1992).

2.1 Leksikografija

Čeprav je že Samuel Johnson pri pisanju prvega angleškega slovarja navajal primere iz literature, je zajemanje in opisovanje besedišča s pomočjo računalniškega korpusa neprimerno lažje in hitrejše. Izpis konkordanc, se pravi korpusnih pojavitev izbrane besedne oblike, in njihovo urejanje po levih ali desnih kolokatorjih nam pokaže frazeološko obnašanje iztočnice, različni sobesedilni vzorci nam pomagajo razbrati njene pomene in rabo. S statističnimi cenilkami, kot sta na primer vzajemna vrednost (MI) ali logaritem razmerij verjetja (LL), pridobimo še podrobnejše podatke o kolokatorjih. Če je korpus oblikoskladenjsko označen, lahko okolje iztočnice raziskujemo še bolj usmerjeno, na primer tako, da si ob samostalniškem geslu prikažemo vse pridevnike ali predložne zveze. Pogostost pojavitve, seveda ob upoštevanju sestave korpusa, njegove (ne)uravnoteženosti in (ne)reprezentativnosti, predstavlja pomemben kriterij pri izdelavi samega geslovnika in pri opisovanju določenega gesla. Pogostejši pomeni naj bi bili navedeni najprej, prav tako nam pogostost pomaga pri opisovanju idiomatike in frazeologije. Spremljevalni korpusi, ki se stalno dopolnjujejo z novimi besedili, so za leksikografa vir podatkov o tem, kaj je v jeziku novo, kaj se spreminja in kaj izumira. In ker korpusi vsebujejo tudi podatke o avtorju, načinu, kraju in času izdaje, lahko opazujemo zvrstno, regionalno ali časovno specifičnost izbranega leksema.

S korpusi se je spremenil tudi način oblikovanja razlage ali definicije gesla. Medtem ko so bile razlage včasih oblikovane po načelu enačbe in (skladenjske) izmenljivosti, npr. *nalupiti – z lupljenjem priti do določene količine česa* (SSKJ), je Cobuildov pristop prinesel razlage v obliki vezanih in lahko berljivih povedi (Pearson 1998), npr. *Če lupimo toliko časa, da imamo olupljenega dovolj, smo nekaj nalupili.* Primeri rabe, ki so jih pri klasičnem pristopu leksikografi skovali sami, so pri Cobuildovem pristopu avtentični stavki iz korpusa.

2.2 Slovnica

Za opazovanje oblikoslovnih in skladenjskih vzorcev v jeziku potrebujemo označeni korpus, avtomatska skladenjska razčlemba (parsing) pa za večino jezikov še ni na voljo. Kljub temu se s polavtomatskimi metodami pospešeno ustvarjajo »globoko« označeni korpusi oziroma »drevesnice« (treebanks), na primer Penn Treebank z 1,6 milijoni besed ali International Corpus of English s približno milijonom besed oziroma 90.000 drevesi. Podobni projekti so na voljo še za bolgarščino, češčino, italijanščino, kitajščino, nemščino, španščino in druge jezike, slovenske »drevesnice« zaenkrat nimamo.[1] Prednost skladenjsko označenih korpusov je, da omogočajo opazovanje in primerjavo slovničnih struktur, predvsem pa sklepanje o tem, kaj je v jeziku bolj in kaj manj tipično. A tudi korpusi brez globinske razčlembe lahko služijo kot pomemben vir za slovnični opis jezika, še posebej glede na zvrst, register in prenosnik. Tu bi si za slovenščino želeli še razširitve korpusa FIDA z na primer govornim korpusom, korpusom internetnih besedil in dopolnitvijo manjkajočih strokovnih področij.

Računalniška obdelava velikih besedilnih zbirk pa nam omogoča učinkovitejše zbiranje podatkov tudi za bolj specifične jezikoslovne raziskave. Kenny (2001) na primer je s pomočjo nemško-angleškega korpusa izvirnih in prevodnih literarnih besedil opazovala strategije pri prevajanju jezikovno kreativnih elementov, kot so novotvorjenke ali slengovski izrazi. Njena hipoteza je predvidevala, da pri prevajanju pogosto pride do normalizacije ali ublažitve izstopajočega pojava. Seveda se pojavi vprašanje, kako s pomočjo korpusa raziskovati izvirnost. Kenny uporabi več metod, med njimi na primer opazovanje enopojavnic ali hapax legomena, ki kažejo na jezikovno neobičajnost, pa tudi tistih besed, ki se pojavljajo zgolj pri določenem avtorju ali v okviru določenega dela. Primerja tudi leksikalno gostoto izvirnikov in prevodov ter ugotavlja, da do normalizacije res prihaja, po drugi strani pa odkrije tudi številne primere, ko prevajalci ublažitev izvirnega jezikovnega sredstva kompenzirajo z nadomestnimi strategijami. Študija vsekakor pokaže, da je z iznajdljivo uporabo korpusnih metod mogoče formalizirati tudi navidez tako neulovljive pojave, kot je izvirnost.

3 Korpusi in uporabniške jezikovne tehnologije

Pomen jezikovnih virov navadno najglasneje poudarjajo jezikoslovci, ki želijo z njihovo pomočjo priti do čim natančnejših opisov jezika, in pa računalniški jezikoslovci, ki se ukvarjajo z razvijanjem metod za avtomatsko obdelavo in analizo naravnega jezika. Z uporabniškega vidika se prizadevanja prvih kažejo v obliki boljših in drugačnih slovarjev, slovnic in drugih jezikovnih priročnikov, pa tudi na primer skozi boljše metode za učenje in poučevanje jezika. Dosežki računalniškega jezikoslovja so širši javnosti manj znani, čeprav se vsaj z nekaterimi jezikovnimi tehnologijami dnevno srečuje.

---

[1] Dober pregled korpusnih virov, tudi označenih, je na strani http://devoted.to/corpora.

3.1 Splošno znana jezikovna orodja

Kot del pisarniškega paketa MS Office imamo za slovenščino na voljo preverjanje črkovanja in slovar sopomenk, obstajajo pa tudi prosto dostopni črkovalniki za urejevalnik Emacs in okolje Unix/Linux. Črkovalnik deluje na podlagi leksikona besed in besednih oblik, ki ga najlaže pridobimo iz korpusa. Naprednejši urejevalniki besedil samodejno zaznajo jezik dokumenta in vključijo ustrezna jezikovna orodja.

Večina mobilnih telefonov je opremljenih s samodejnim dopolnjevanjem besed pri pisanju kratkih sporočil, ki prav tako temelji na slovarju pogostih besed in besednih oblik. Za slovenščino je to orodje še precej nepopolno, na splošno pa se mora tak dopolnjevalnik ravnati po pogostostih rabe besed v kratkih sporočilih, saj je ta besedilna vrsta – podobno kot e-poštna sporočila – zelo specifična in nam podatki o rabi in pogostosti iz referenčnega korpusa le malo koristijo.

Dobro znana, čeprav neprimerno bolj zapletena aplikacija je tudi strojni prevajalnik, ki ga za jezikovni par angleščina-slovenščina razvija podjetje Amebis d. o. o. Čeprav tehnologija strojnega prevajanja tradicionalno ni temeljila na korpusnih virih, ampak na obsežnih dvojezičnih leksikonih in transformacijskih pravilih, se v zadnjih petnajstih letih razvoj osredotoča na statistične metode in strojno učenje na podlagi vzporednih korpusov. Eden takih sistemov je Egypt (Och in Ney 2000), prednost statističnega prevajanja pa je jezikovna neodvisnost, saj lahko sistem naučimo na vzporednem korpusu katerega koli jezikovnega para. Za komercialne namene statistično strojno prevajanje sicer še ni zrelo, so pa v teku tudi poskusi učenja sistema Egypt na slovensko-angleških vzporednih korpusih IJS-ELAN in TRANS (Vičič in Erjavec 2002).

Preden zapustimo jezikovne tehnologije za slovenščino in pogledamo naokrog, je treba omeniti pri nas zelo živahno področje govornih tehnologij, ki se uspešno vključujejo v različne uporabniške aplikacije, na primer govorne vmesnike pri mobilni telefoniji (poslušanje e-pošte in kratkih sporočil), pripomočke za slabovidne in slepe, učne pripomočke za otroke in še marsikaj. Ker se govornim tehnologijam posveča drug prispevek v tej publikaciji, se tu z njimi ne ukvarjamo podrobneje.

3.2 Pomensko usmerjene aplikacije

Z uporabo jezikovnih tehnologij so se razvila številna orodja, ki skušajo na tak ali drugačen način poleg oblikoskladenjskih značilnosti naravnega jezika zajeti tudi pomen. Sem sodi na primer samodejno razvrščanje dokumentov (Document Classification), pri katerem mora sistem prepoznati ključne besede, jih razvrstiti po pomembnosti in na podlagi tega dokument razvrstiti v eno od danih kategorij. Prepoznavanje ključnih besed je lahko statistično, se pravi s primerjavo pogostosti besed v celotni zbirki dokumentov in po posameznih dokumentih, lahko pa temelji na splošnih tezavrih tipa WordNet ali področnih ontologijah. Samodejno razvrščanje je pomembno na primer pri spletnih iskalnikih in imenikih, v podjetjih in ustanovah, ki se srečujejo z velikim pritokom dokumentov, v dokumentalistiki in bibliotekarstvu in drugod. Sorodne aplikacije so se razvile tudi za upravljanje z

elektronsko pošto. Program (npr. Xtramindov Mail-Minder; http://www.xtramind.com/english/html/products/email_response_management.html) na podlagi analize prejetih sporočil predlaga najučinkovitejšo razdelitev v mape, nato pa pri vsakem prihajajočem sporočilu samodejno zazna jezik in ga s pomočjo ključnih besed, značilnih vzorcev in podatkov iz oglavja sporočila razvrsti v eno od map. Dodatne možnosti programa vključujejo še samodejno povzemanje sporočil in samodejno tvorjenje odgovorov.

Povzemanje (Text Summarization) je v času vsesplošne informacijske prezasičenosti nedvomno koristna aplikacija. Danes je na voljo prek deset komercialnih povzemovalnikov,[2] med njimi tudi Microsoftov, ki ga je mogoče vključiti v Word in druge programe. Zgodnji povzemovalniki so prav tako temeljili na luščenju ključnih besed in vrednotenju povedi glede na informativno težo. Povzetek, katerega dolžino lahko uporabnik določi sam, je tako sestavljen iz ustreznega števila visoko uvrščenih povedi, katerih notranja zgradba ostane nespremenjena. Naprednejša orodja vključujejo jezikovno odvisne komponente oblikoskladenjske analize in zmorejo mnogo več, med drugim tudi preoblikovanje povedi tako, da je zajeta le ključna informacija, pa tudi povzemanje množice dokumentov.

Pomensko usmerjene aplikacije se srečujejo s specifičnimi jezikovnotehnološkimi problemi, ki ostajajo predmet živahnih raziskav. Eden izmed njih je razdvoumljanje besed (Word Sense Disambiguation), kjer skušamo za večpomensko besedo v danem besedilu na podlagi njene okolice ugotoviti, kateri od možnih pomenov je zares v igri. V ta namen se dobro obnesejo statistične metode, kot so skupkanje (clustering) ali razvrščanje (classification), kjer sobesedilo večpomenske besede predstavlja njen kontekstni vektor. Z izračunom razdalj med posameznimi vektorji je mogoče ugotoviti, za kateri pomen gre. Zanimiv prikaz eno- in dvojezičnega skupkanja z vizualizacijo rezultatov je bil zgrajen na inštitutu CSLI Univerze v Stanfordu v okviru projekta Infomap <http://infomap.stanford.edu>.

Za razvoj vseh naprednejših jezikovnih orodij so neobhodno potrebni jezikovni viri, kot so korpusi, oblikoslovni leksikoni, splošni in področni pomenski tezavri, leksikoni lastnih imen in drugi. Šele z njimi je namreč mogoče razviti jezikovno specifična orodja, ki so zmožna pomenske obdelave besedil.

4 Iskanje podatkov in rudarjenje besedil

Pomembno področje, na katerem se jezikovne tehnologije šele uveljavljajo, je iskanje podatkov (Information Retrieval). Izraz pomeni dostopanje do relevantnih dokumentov v velikih zbirkah na podlagi poizvedbe v naravnem jeziku, se pravi spiska besed, iskane fraze ali cele povedi v obliki vprašanja. Najbolj znana in največja zbirka dokumentov, do katerih dostopamo na tak način, je svetovni splet, na številnih strokovnih področjih pa se vzdržujejo tudi drugi elektronski arhivi, na

[2] Pregled jezikovnih tehnologij, vključno s povzemovalniki, nudi stran Language Technology World - http://www.lt-world.org.

primer zbirke pravnih aktov, medicinskih člankov, tehničnih opisov proizvodov, upravnih kartotek itd. Najenostavnejši iskalniki v besednem kazalu preprosto poiščejo dokumente, ki vsebujejo iskane besede, in jih razvrstijo glede na število njihovih pojavitev. A to zagotovo ni najučinkovitejši način, kajti pogostost besede ni vselej merilo za njeno ključnost, poleg tega pa je s tem iskanje omejeno le na dano besedno obliko.

Za merjenje relevantnosti besede za določeni dokument znotraj zbirke se na splošno uporablja cenilka tf.idf (Term Frequency – Inverse Document Frequency) (Baeza-Yates in Ribeiro-Neto 1999), ki temelji na predpostavki, da je beseda tem bolj značilna za posamezni dokument, čim manj ostalih dokumentov jo vsebuje, in čim večkrat se v tem dokumentu pojavlja. A za uspešnejše iskanje je treba, še posebej pri oblikoslovno razgibanih jezikih, vključiti vse besedne oblike, po možnosti pa še njene sopomenke. Tako danes mnogi spletni iskalniki vključujejo jezikovna orodja, ki najprej samodejno razpoznajo jezik poizvedbe, nato pa izločijo nepomembne besede, kot so vezniki, členi in predlogi, razen če je niz označen kot fraza. Zatem lahko sledi krnjenje, ki je poenostavljena različica lematizacije, ali prava lematizacija oziroma širjenje poizvedbe z drugimi besednimi oblikami. Mnogi spletni iskalniki prepoznajo tudi napačno zapisane besede, in sicer preprosto na podlagi pogostosti besednih oblik.

Naprednejše iskanje, ki se uporablja večinoma za področne zbirke dokumentov, vključuje širjenje iskalnega pogoja s podobnimi besedami ali sinonimi iz tezavra. Namesto ročno izdelanih tezavrov se navadno še bolje obnesejo avtomatski, ki iz dane zbirke dokumentov izluščijo skupinice besed s podobnimi kontekstnimi vektorji. Posebno področje iskanja podatkov pa je medjezično iskanje (Cross-Language Information Retrieval), pri katerem iskalnik za dano poizvedbo poišče tudi dokumente v jezikih, ki niso enaki kot jezik poizvedbe (glej tudi Dimec 2002). Medjezično iskanje lahko temelji na strojnem prevajanju, kjer se prevede bodisi zgolj poizvedba bodisi celotna zbirka dokumentov. Na področjih, kjer so na razpolago obširne večjezikovne ontologije, kot je na primer Unified Medical Language System (UMLS) za medicino, lahko iskanje poteka tudi preko t. i. pojmovnega prenosa. V poizvedbi se najprej poiščejo pomembni področni pojmi, ki so označeni z jezikovno neodvisno kodo, prek te kode pa je mogoč prenos poizvedbe v ključne izraze ciljnega jezika. Prototip takega sistema je bil razvit v projektu MuchMore in je dostopen na spletu na naslovu http://lit.dfki.uni-sb.de:8000/prototype/index.html. Za področja, kjer takšnih ontologij ni, je s pomočjo dvojezičnega slovarja in dane (dvojezične) zbirke dokumentov mogoče izdelati dvojezični tezaver pomensko podobnih besed (similarity thesaurus), ki prav tako omogoča učinkovito prevajanje in širjenje poizvedbe.

Nekoliko sorodno področje, kjer se računalniško jezikoslovje srečuje z informatiko in umetno inteligenco, je rudarjenje besedil (Text Mining) kot posebno področje rudarjenja podatkov. Tu ne gre za dostopanje do dokumentov, ampak za pridobivanje znanja iz besedilnih zbirk, še posebej takšnega znanja, ki prej ni bilo eksplicitno dostopno.[3] Metode rudarjenja besedil obsegajo luščenje terminologije in ključ-

[3] Povzeto po študijskih gradivih Hinricha Schuetzeja in Chrisa Manninga za predmet Text Mining na Stanfordski univerzi, http://www-csli.stanford.edu/~schuetze/.

nih besed na eni strani, kar v napredni obliki postane luščenje ontološkega znanja, se pravi pojmov in razmerij med njimi, na drugi strani pa gre za odkrivanje novih povezav in korelacij med podatki, pogosto s pomočjo vizualizacije, kar vodi do novih strokovnih spoznanj.

Gradnja ontologij oziroma jezikovno neodvisnih mrež pojmov, njihovih lastnosti in razmerij med njimi je v zadnjem času posebej aktualno področje, ki je v ospredje stopilo s semantičnim spletom, idejo o pomensko organiziranem internetu, kjer so spletne informacijske in druge storitve dostopne prek metapodatkov, standardiziranih ontologij in pametnih agentov <http://www.w3.org/2001/sw/>. Ker za mnoga področja obširnih ontologij še ni, njihova ročna izdelava pa zahteva precej truda in časa, se pospešeno razvijajo metode avtomatske gradnje ontologij iz (predvsem) besedilnih virov (glej npr. Maedche in Staab 2001).

V naslednjem razdelku nekoliko pobliže opišemo dva primera izrabe jezikovnih virov, od katerih je prvi neposredno, drugi pa posredno povezan tudi s postopki rudarjenja besedil.

5 Dva primera izrabe jezikovnih virov

5.1 Pridobivanje novih medicinskih spoznanj

Medicina je izrazito kompleksno in široko področje, kjer se neprestano pojavljajo odkritja novih bolezni, povzročiteljev, učinkovin in interakcij med njimi. Najbolj utečena pot za širjenje novih spoznanj je preko strokovnih člankov, ki izhajajo v nekaj tisoč mednarodnih serijskih publikacijah in zbornikih znanstvenih srečanj. Danes je precejšnji del teh publikacij dostopen tudi preko interneta, pri čemer je najbogatejši vir baza Medline oziroma njen javno dostopni vmesnik PubMed <http://www.ncbi.nlm.nih.gov/PubMed/>, ki vsebuje bibliografske podatke člankov z naslovi in povzetki iz več kot 4.000 strokovnih revij, skupaj preko 10 milijonov člankov od leta 1966 do danes. Ni torej presenetljivo, da se je pojavila zamisel o izkoriščanju tega obsežnega korpusa za odkrivanje novih povezav in znanstvenih spoznanj.

Naloga se na prvi pogled morda zdi nesmiselna, kajti če želimo iz korpusa izluščiti določeno znanje, mora biti to na tak ali drugačen način ubesedeno, to pa pomeni, da ni novo. Toda medicina je tako razvejano področje, da so mnoga znana dejstva o, denimo, povezavi med določeno farmakološko snovjo in odzivom organizma omejena zgolj na tisto specialistično stroko, ki se s tem ukvarja. Obenem so o istem odzivu organizma pod drugačnim specialističnim vidikom morda znana druga dejstva, povezava teh dejstev pa morda pomeni novo medicinsko spoznanje.

Takšen pristop prvi opisuje Swanson (1991), ki s pomočjo naslovov in povzetkov člankov iz Medlinea skuša ugotoviti vzročne povezave med simptomi, zdravili in rezultati. V eni svojih raziskav se je osredotočil na migrenski glavobol in prehranske vzroke zanj. Njegova metoda je razmeroma enostavna: Če imamo problem A (migrenski glavobol) in ciljni prostor vzrokov C (prehrana), lahko zberemo litera-

turo o A in literaturo o C. Če obstajajo takšni elementi B, ki se v literaturi pogosto pojavljajo tako v zvezi z A kot v zvezi s C, lahko postavimo hipotezo A → B → C. Pri migrenskem glavobolu se je od prehranskih vzrokov osredotočil na magnezij in v Medlineu zasledil naslove, kot so:

- *Stress is associated with migraines* [Stres je povezan z migrenami]
- *Stress can lead to a loss of magnesium* [Stres lahko povzroči pomanjkanje magnezija]
- *Calcium channel blockers prevent some migraines* [Zaviralci kalcijevih kanalčkov lahko preprečujejo migreno]
- *Magnesium is a natural calcium channel blocker* [Magnezij je naravni zaviralec kalcijevih kanalčkov]

Presek dveh specialističnih področij torej pomaga identificirati faktorje, ki pripeljejo do novih povezav, v tem primeru do povezave med pomanjkanjem magnezija in migrenskim glavobolom. Pri zgoraj navedenih primerih sta vmesna člena *stres* in *zaviralci kalcijevih kanalčkov*. Swanson je na ta način postavil več novih hipotez in nekatere tudi objavil v medicinskih revijah, nekaj pa so jih kasneje tudi potrdili z eksperimentalnimi dokazi.

Ta zgodnji pristop je temeljil na precej rudimentarnih postopkih in je zahteval veliko človeškega dela pa tudi medicinskega znanja. Odtlej so se razvila naprednejša orodja, ki uporabljajo jezikovno analizo besedil, semantično označevanje medicinskih terminov in pomenskih razredov, razvrščanje dokumentov po pomembnosti in druge metode za čim večjo avtomatizacijo procesa odkrivanja znanja. Tako orodje je sistem DAD (Weeber in soavtorji 2000), ki medicinskemu strokovnjaku omogoča razvijanje in testiranje hipotez na podlagi baze člankov PubMed in medicinskega metatezavra UMLS <http://www.nlm.nih.gov/research/umls/>. Z uporabniško prijazno zasnovo, ki v ozadju kljub temu skriva napredne jezikovne in podatkovne obdelave, je njegovim avtorjem uspelo ustvariti inovativen pripomoček za rudarjenje besedil.

Sicer pa se rudarjenje podatkov v biomedicini zadnje čase usmerja predvsem v genske raziskave, ki ne temeljijo toliko na rudarjenju besedil kot na odkrivanju vzorcev v – doslej znanih – genomih. Zaradi izredne vplivnosti, še posebej pa finančne vrednosti teh raziskav se na to temo po svetu letno zgodi kar nekaj znanstvenih posvetov, eden od pomembnejših je sklican tudi za jesen 2003 v »sosednjem« Dubrovniku.[4]

[4] Data and Text Mining for Bioinformatics, ECML/PKDD 2003, http://www.cs.kuleuven.ac.be/conference/ecmlpkdd/.

5.2 Dvojezično luščenje terminologije

S preprostimi besedami bi to tehnologijo lahko opisali kot (pol)avtomatsko strokovno slovaropisje, vendar namen avtomatskega luščenja terminologije iz besedil – v nasprotju z razširjenim prepričanjem – ni samodejna izdelava slovarjev, temveč podpora različnim dejavnostim in tehnologijam, kjer je pisanje strokovnih slovarjev le ena izmed njih. Dvojezično luščenje terminologije (Bilingual Term Extraction) (prim. Vintar 2002) pomeni iskanje strokovnih izrazov in njihovih prevodnih ustreznic v dvojezičnem, navadno vzporednem korpusu strokovnih besedil. Potrebe po dvojezičnih terminoloških virih imajo predvsem prevajalci, poleg njih pa tudi terminologi, tehnični pisci, dokumentalisti in bibliotekarji ter sami strokovnjaki določenega področja. Samodejno razpoznavanje terminologije je tudi komponenta aplikacij, kot so prevajalska namizja, iskalniki podatkov, povzemovalniki in drugo.

Pri luščenju terminologije za prevajalske namene ima korpusni pristop še to prednost, da s samo sestavo korpusa lahko bistveno vplivamo na kakovost, sodobnost in doslednost dobljenih rezultatov, hkrati pa tako pridobljeni izrazi in njihove ustreznice predstavljajo dinamičen in po meri zgrajen terminološki vir. Težava, da za mnoga področja nimamo na voljo ustreznih vzporednih korpusov, je sicer še vedno tu, vendar se z vse bolj razširjeno uporabo pomnilnikov prevodov v prevajalskih okoljih postopoma zmanjšuje.

Sama tehnologija je sestavljena iz iskanja izrazov v enem in drugem jeziku ter iskanja prevodnih ustreznic. Za ugotavljanje, kaj je v strokovnem korpusu termin in kaj ne, uporabljamo različne statistične in jezikoslovno utemeljene postopke, ki jih lahko povzamemo z naslednjimi hipotezami in odgovori nanje:

- Če je izraz termin, se bo v strokovnem besedilu pojavljal bolj pogosto kot v splošnem besedilu. Na podlagi te hipoteze merimo ključnost besed, in sicer s primerjavo pogostosti v strokovnem korpusu in referenčnem korpusu splošnega jezika.
- Mnogi termini vsebujejo tujejezične sestavine, simbole ali kratice, ki jih v splošnih besedilih ne srečamo. Pri iskanju terminov se torej osredotočimo na elemente, ki so neobičajni oziroma jih leksikon opazovanega jezika ne vsebuje.
- Če je večbesedna enota termin, se bo v bolj ali manj nespremenjeni obliki pojavljala skozi celoten strokovni korpus. Osredotočimo se torej na kolokacije, ki so stabilne.
- Termini imajo tipične skladenjske oblike, npr. pridevnik + samostalnik (*natrijev klorid*). Ti skladenjski vzorci so deloma vezani na splošne slovnične značilnosti jezika, v veliki meri pa so odvisni tudi od strokovnega področja. Ugotavljanje terminološko relevantnih vzorcev je zato pomemben del luščenja terminologije, ki ključno vpliva na kakovost rezultatov. Na vsak način pri tem potrebujemo oblikoskladenjsko označeni korpus.

S pomočjo teh načel in kombinacije statističnih in jezikovnih orodij je mogoče iz korpusa izluščiti terminološke kandidate. Za iskanje prevodnih ustreznic se uporablja postopek besedne poravnave, pri katerem za vsako besedno obliko v prvem

jeziku korpusa izračunamo najverjetnejše prevodne ustreznice v drugem jeziku. Pri oblikoslovno bogatih jezikih se obrestuje, če pred luščenjem prevodnih parov opravimo lematizacijo, se pravi pretvorbo v osnovno obliko. Na voljo je nekaj prosto dostopnih programov za besedno poravnavo, med njimi Twente (Hiemstra 1998) in Giza++, ki se je razvila v okviru prej omenjenega projekta Egypt. Ko smo iz korpusa pridobili dvojezični leksikon, je prej prepoznane termine mogoče poravnati z njihovimi prevodnimi ustreznicami.

Orodja za dvojezično luščenje terminologije niso novost, a si kljub temu pot v komercialne aplikacije šele utirajo. Novejše različice prevajalskih orodij, kot sta TRADOS in DéjàVu, vključujejo preproste prepoznavalnike izrazov in njihovih ustreznic, boljši sistemi pa so še vedno bodisi skriti v velikih korporacijah kot interno programje bodisi v razvoju na raziskovalnih ustanovah.

Prvi poskus takega sistema je bil zgrajen tudi za jezikovni par angleščina-slovenščina (Vintar 2002). Za testna korpusa smo uporabili vzporedna besedila s področij jedrske tehnike in gospodarske zakonodaje. Kakovost izluščenih terminoloških parov smo ocenjevali s pomočjo področnega strokovnjaka in poklicne terminologinje. Še posebej slednja je rezultate ocenila razmeroma ugodno, saj je kakovost predlaganih izrazov presegala 65 odstotkov, kakovost prevodnih ustreznic pa je znašala celo 78 odstotkov.

Ker so takšni sistemi najbolj smiselni za uporabo v večjih prevajalskih okoljih, kjer nastajajo velike količine vzporednih besedil, je tudi njihovo zasnovo treba ustrezno prilagoditi uporabniškim potrebam. Iz tega razloga ni pričakovati, da bi se v bližnji prihodnosti pojavili luščilniki izrazja za širše uporabniške kroge, zato pa lahko uporaba specifično zasnovanih orodij bistveno poveča učinkovitost podjetij, ki jim je ukvarjanje s strokovnim izrazjem del vsakodnevnega posla.

6 Sklep

Jezikovna infrastruktura, kamor sodijo korpusi, leksikoni, tezavri, označevalniki in druga orodja za obdelavo besedil ter široka paleta govornih tehnologij, nikakor ne služi le jezikoslovcem, ampak predstavlja bistveni del sodobne informacijske infrastrukture. Globalni delež jezikovnih podatkov v razmerju do numeričnih in drugih strukturiranih podatkovnih virov je po večini ocen med 70 in 80 odstotkov v prid prvih, zato jezikovne tehnologije odpirajo vrata do večine računalniško berljivega človeškega znanja.

Računalniško jezikoslovje je v zadnjih nekaj desetletjih doživelo revolucionarne premike. Od slovničnih pravil je preko statističnih metod nastopila doba strojnega učenja, od skromnih podatkovnih virov in polžje obdelave smo danes z elektronskimi viri besedil in procesorskimi zmogljivostmi tako rekoč neomejeni. Računalniki prevajajo, govorijo, urejajo prihajajočo pošto in nanjo odgovarjajo, v nekaj sekundah izmed milijona dokumentov izberejo tistega, ki ga iščemo, nadzorujejo vsebino našega hladilnika in manjkajoče predlagajo v obliki nakupovalnega spiska. Pa vendar smo ob misli na to, da bi računalniki naravni jezik obdelovali podobno uspešno kot druge vrste podatkov, še vedno na pragu prej omenjenih vrat. Ko jih

bomo zares in nepreklicno prestopili, bo verjetno treba računalnik preimenovati v, na primer, jezikalnik…

Literatura

Baeza-Yates, R. in Ribeiro-Neto, B., 1999: *Modern Information Retrieval.* Boston: Addison Wesley Longman.

Dimec, J., 2002: Medjezično iskanje. *Knjižnica* 1–2. Ljubljana.

Hiemstra, Djoerd, 1998: Multilingual Domain Modelling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. *Proceedings of the 8th CLIN meeting*. 41–58.

Kenny, D., 2001: *Lexis and creativity in translation. A corpus-based study.* Manchester: St. Jerome.

Leech, G. N., 1991: The state of the art in corpus linguistics. Aijmer, K., Altenberg, B. (ur.): *English Corpus Linguistics*. London: Longman. 8–29.

Maedche, A. in Staab, S., 2001: Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2). 72–79.

McEnery, T. in Wilson, A., 1992: *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Och, F.-J. in Ney, H., 2000: Statistical machine translation. *Zbornik rednega letnega srečanja European Association for Machine Translation (EAMT 2000)*. Ljubljana. 39–46.

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

Swanson, D. R., 1991: Analysis of Unintended Connections Between Disjoint Science Literatures. *Proceedings of SIGIR 1991*. 280–289.

Thomas, J. in Short, M., 1996: *Using Corpora for Language Research*. London: Longman.

Vičič, J. in Erjavec, T., 2002: Vsak začetek je težak: avtomatsko učenje prevajanja slovenščine v angleščino. *Zbornik konference Jezikovne tehnologije, v sklopu konference Informacijska družba 2002*. Ljubljana: Institut Jožef Stefan.

Vintar, Š., 2002: Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil. *Zbornik konference Jezikovne tehnologije, v sklopu konference Informacijska družba 2002*. Ljubljana: Insitut Jožef Stefan.

Weeber, M., H. Klein, A. R. Aronson, J. G. Mork, L. Jong-van den Berg in R. Vos, 2000: Text-Based Discovery in Biomedicine: The Architecture of the DAD-system. *Proceedings of the American Medical Informatics Association 2000 Symposium, Los Angeles, CA*.

Spletne strani

Bookmarks for Corpus-Based Linguists <http://devoted.to/corpora>
Concept Based Information Representation and Retrieval – Infomap <http://infomap.stanford.edu>
Data and Text Mining for Bioinformatics, ECML/PKDD 2003 <http://www.cs.kuleuven.ac.be/conference/ecmlpkdd/>

FIDA – korpus slovenskega jezika <http://www.fida.net>
Hinrich Schuetze – Text Mining (CSLI Stanford) <http://www-csli.stanford.edu/~schuetze/>
Language Technology World <http://www.lt-world.org>
Multilingual Concept Hierarchies for Medical Information Retrieval and Organization – MuchMore Demo <http://lit.dfki.uni-sb.de:8000/prototype/index.html>
PubMed <http://www.ncbi.nlm.nih.gov/PubMed/>
Unified Medical Language System <http://www.nlm.nih.gov/research/umls/>
W3C – Semantic Web < http://www.w3.org/2001/sw/>
Xtramind MailMinder
<http://www.xtramind.com/english/html/products/email_response_management.html>