Uporaba korpusa pri urejanju spletnega terminološkega slovarja

Katarina Puc,[*] Tomaž Erjavec[‡]
*Ljubljana, katarina.puc@drustvo-informatika.si
‡Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

V članku opišemo, kako se pri urejanju spletnega terminološkega slovarja Islovar uporablja korpus informatike, ki se oblikuje iz zbornikov posvetovanja Dnevi slovenske informatike. Islovar vsebuje izrazje s področja informatike in je odprt za uporabo, pa tudi za prispevke uporabnikov. Zbranih izrazov je veliko več kot urejenih, ker je končno urejanje zahteven postopek. Pri urejanju uredništvo upošteva vse ugotovljene sinonime izrazov, ki jih hkrati ovrednoti glede na uporabo. Korpus informatike je odličen vir, ker zajema prispevke številnih, različnih avtorjev. V prispevku predstavimo izdelavo in značilnosti korpusa, podamo primere uporabe korpusa ter drugih elektronskih virov in zaključimo z načrti za prihodnost.

Using a corpus for editing an on-line terminological dictionary

In the paper the use of DSI corpus for editing the on-line terminological dictionary Islovar is described. DSI is a specialized corpus, created from proceedings of Slovene Informatics Conferences. The Islovar dictionary includes terms relating to the field of informatics, with open access for users and also for contributors of new words and commentaries. Because compiling the final edition of the dictionary is a demanding process, the dictionary still contains more collected than edited entries. The editors take into consideration all the discovered synonyms which are evaluated and labelled according the usage. Originating from articles by many authors, the corpus is an excellent source of informatics terms. In the paper, the creation and the characteristics of the corpus are described. Some examples of use in editing Islovar and a comparison with reference corpora and other electronic sources are given. Finally, we mention some plans for the future.

1. Uvod

Islovar, http://www.islovar.org/, je slovenski spletni terminološki slovar informatike, ki ga ureja Slovensko društvo INFORMATIKA od aprila 2001. Slovar navaja tudi angleške ustreznice, tako da lahko iščejo uporabniki slovenske izraze tudi iz angleških, urejeni slovarski sestavki pa vsebujejo slovensko razlago in kvalifikatorje.

Ker so njegovi avtorji informatiki, je naravno, da se pri uporabi in pri urejanju tega slovarja izkoriščajo vse prednosti informacijskih tehnologij. Prav pri izdajanju slovarjev so se te prednosti v zadnjih letih posebej izkazale, o čemer priča veliko število spletnih slovarjev v vseh jezikih. Posebnost Islovarja je v njegovi odprtosti ne samo za branje, temveč tudi za zbiranje in urejanje.Uporabniki slovarja lahko nove izraze prispevajo, dodajajo razlage in obstoječe sestavke komentirajo.

Odprtost, značilna za objave na svetovnem spletu, je velika prednost, ker omogoča sodelovanje velike populacije. Po drugi strani pa je lahko tudi problem: ker lahko vsak objavlja karkoli, sta kakovost in zanesljivost spletnih dokumentov pogosto vprašljivi. Uredništvo Islovarja to rešuje z dogovorjenim uredniškim postopkom in tako zagotavlja, da so pri urejanju vsi zapisi v Islovar večkrat pregledani.

Namen Islovarja je poenotenje informacijskega izrazja in opremljanje novih pojmov s splošno privzetimi slovenskimi ustreznicami. Zato uredništvo deluje v skladu z načeli upoštevanja zanesljivosti, kakovosti in ažurnosti vsebine. Islovar je informativni in normativni slovar, beleži vse pojave izraza, jih ovrednoti in ustrezno označi.

Za zbiranje in analiziranje izrazja se je v slovaropisju uveljavila jezikovna tehnologija korpusov. Za raziskave specializiranih besedil nastajajo specializirani korpusi, ki so navadno manjši in vsebujejo jezik v točno določeni rabi. Pogosto jih uporabnik izdela sam, s točno določenim namenom (Arhar, 2006).

Tak specializirani korpus je korpus informatike (v nadaljevanju korpus DSI), ki vsebuje članke iz zbornikov posvetovanja Dnevi slovenske informatike iz let 2003−2006. Ta posvetovanja so letna in množično obiskana, obravnavajo pa različno tematiko, od strateških vidikov informatike, do operacijskih raziskav. Prispevki avtorjev se objavljajo v zbornikih. Tematika je pretežno aktualna in se menja iz leta v leto, avtorji besedil pa so informatiki iz prakse in z univerz. Prispevki so lektorirani, tako da žargonski izrazi zvečine niso zajeti.

V članku bomo opisali uredniški postopek v Islovarju, zgradbo korpusa DSI in kako ta korpus uporabljamo pri slovaropisju.

2. Uredniški postopek v Islovarju

Islovar je spletni računalniški program. Od novembra 2004 deluje že v drugi izdaji, ki je vnesla mnoge izboljšave v uporabniškem in v uredniškem vmesniku. Zlasti je bil izboljšan iskalnik, ki omogoča iskanje tudi po podobnih izrazih in zato uporabnikom dovoljuje tudi manjše napake pri vnosu iskanih iztrazov.

Uredniški vmesnik ima številne nove možnosti: iskanje po Islovarju po raznih kriterijih, tudi po avtorjih sestavkov, vpogled v novo zapisane sestavke in v zgodovino sprememb, delo z zbirkami.

Delo urednikov poteka neposredno v spletnem slovarju po dogovorjenem uredniškem postopku. Slovarski sestavki imajo posebno oznako, ki kaže na stopnjo obdelave sestavka in zanesljivost vsebine. Oznako *predlog* prejmejo izrazi takoj po vnosu v slovar, po prvem pregledu, ko se preverja vsebinska primernost za področje informatike, pa so izrazi označeni kot *pregledani*.

Naslednja značilnost uredniškega postopka je analiza zbirke izrazov, ki jo pripravi urednik ali strokovna skupina. Zbirka je vsebinsko povezana družina, ki zajema tudi vse izraze, ki so bili uporabljeni v razlagah v tej zbirki. Po javni razpravi o predlagani zbirki, ki se je udeležujejo vsi uredniki Islovarja, sledijo končni popravki in opredelitev *strokovno pregledano*.

Do tu se uredniški postopek usmerja zlasti na določitev vsebinskega obsega, strokovne ustreznosti strokovnega izraza in natančnosti razlage. Sledi slovaropisna obravnava, kjer se poskrbi za formalno pravilnost slovarskih sestavkov glede na slovaropisna priporočila in uskladitev z že urejenimi slovarskimi sestavki. Po slovaropisni obravnavi strokovna skupina oziroma urednik, ki je pripravila oziroma pripravil zbirko, ponovno preveri pravilnost vsebine, sledi natančen formalni pregled, nakar se slovarski sestavki označijo kot *urejeni*. Tudi ti sestavki se kasneje lahko še spreminjajo.

Vse te razprave potekajo delno na sestankih, delno neposredno v Islovarju (kot komentarji k sestavkom ali v okviru razprave v forumu) ali po elektronski pošti. Udeleženci na sestankih uporabljajo osebne računalnike z neposredno povezavo na spletno stran Islovarja, tako da je vse delo neposredno zabeleženo in vidno vsem urednikom. Izpis na papirju je nujno potreben šele pri končnem formalnem pregledu.

V tem postopku uredniki poleg izdaj v knjižni obliki izkoriščajo vse vire, ki so dostopni v elektronski obliki na spletu. V slovenščini so na voljo številni slovarčki, glosarji in Leksikon računalništva in informatike, uporabljajo pa se predvsem angleški spletni slovarji in leksikoni ter drugi spletni dokumenti, ki jih najdemo s spletnima iskalnikoma najdi.si in Google.

Ker poimenovanje pojmov na tem področju nikakor ni poenoteno in uporabljajo v slovenščini razni avtorji različne izraze za isti pojem, se mora uredništvo odločati, kako ovrednotiti posamezne ustreznice in ali predlagana razlaga res ustreza sodobni uporabi. Pri tem odločanju so pomembni dostopnost, zanesljivost in ažurnost virov.

Sorazmerno lahko dostopen je podatek o pogostosti uporabe. Zelo priročna pri tem sta iskalnika najdi.si in Google v slovenščini. Rezultati takega iskanja pa so žal pogosto nezanesljivi. Številni pomembni dokumenti na spletu sploh niso objavljeni, nesorazmerno pa so zastopani prispevki raznih posameznikov, razprave v forumih, in seveda objave prodajalcev opreme. Zato je rezultate spletnega iskanja treba temeljito pretehtati. Kot zanesljiv spletni vir obravnava uredništvo seminarska in diplomska dela, objavljena na spletu, pa tudi programe fakultet in reviji Monitor in Moj mikro.

Zanesljiv vir je tudi terminološka zbirka Evroterm, kjer so koristni primeri uporabe in prevodi v druge jezike. Žal pa je obdelana predvsem zakonodaja Evropske unije in zato informatika samo obrobno.

Zaradi ažurnosti, tako pomembne pri informatiki, so samo delno uporabni dokumenti, ki so stari 5 let in več. Zato skoraj ni uporaben referenčni korpus Nova beseda, ki strokovne izraze navaja pretežno iz dnevnika Delo, iz Monitorja pa iz 2000 do leta 2002. Tudi Leksikon (Pahor, 2002) je žal že v marsičem zastarel.

Glede ažurnosti in zanesljivosti je korpus DSI za uporabo uredništva Islovarja daleč najboljši. Nudi nam vpogled v primere uporabe in pogostost uporabe. Ker posvetovanja Dnevi slovenske informatike obravnavajo široko, zlasti aktualno tematiko, je korpus uporaben v velikem številu primerov.

3. Korpus DSI

Ker zborniki pokrivajo isto področje kot slovar, obenem pa so strokovni prispevki dragocen vir svežega slovenskega izrazja, smo se že leta 2003 dogovorili, da se zbornike pretvori v korpus, ki bi nato lahko služil kot podpora pri izdelavi slovarja (Erjavec in Vintar, 2004).

Korpus DSI je bil narejen iz digitalnih originalov (Microsoft Word) konferenčnih prispevkov. Ti so zapisani v skladu s predlogo konference, kar v marsičem olajša nadaljnji postopek pretvorbe. Dokumente smo najprej pretvorili v XML, nato pa s filtrom XSLT ta XML pretvorili v besedilo korpusa. Filter iz dokumentov izloči nebesedilne elemente in tiste razdelke, ki so pisani v angleškem jeziku (angleški povzetek, bibliografija). Tu je potrebno omeniti, da se je v vsakem od zbornikov pojavilo par prispevkov, ki zaradi napak v formatu Word niso bili pretvorjeni; zato je število besedil v korpusu nekaj manjše od števila prispevkov v zbornikih, število besed pa zaradi tega, in zaradi omenjenih izpustov delov besedil tudi manjše kot število besed v zbornikih.

V drugi fazi se besedilo korpusa jezikoslovno označi. To smo storili s pomočjo programa *totale* (Erjavec et al., 2005), ki besedilo naprej tokenizira (razdeli besedilo na besede, ločila in povedi), nato oblikoslovno označi (vsaki besedi pripiše njeno oblikoslovno oznako iz nabora MULTEXT-East za slovenski jezik) in lematizira (določi besedam njihovo osnovno obliko). Program sicer označi vse besede v korpusu, tako znane kot neznane, vendar pa pri označevanju dela tudi napake; največ problemov povzročajo angleške besede in okrajšave.

Korpus trenutno obsega štiri zbornike (2003–2006), velikost korpusa po posameznih letnikih in skupno pa je podana v Tabeli 1.

Letnik	*Besedil*	*Odstavkov*	*Stavkov*	*Besed*
2003	111	3.164	9.791	196.883
2004	109	2.893	9.273	200.287
2005	123	3.546	10.474	223.635
2006	137	4.277	12.022	262.260
Σ	480	13.880	41.560	883.065

Tabela 1. Velikost korpusa DSI 2003−2006

Poglejmo še besedišče korpusa. V Tabeli 2 je podano število različnih lem, torej osnovnih oblik besed po treh najbolj zanimivih besednih vrstah, ter za vse besedne vrste. Kot rečeno, prihaja pri lematizaciji tudi do napak, zato so razmeroma zanesljive samo leme, ki se v korpusi pojavijo večkrat; tabela poda števila za vse leme, ter za tiste, ki se pojavijo vsaj dvakrat oz. trikrat.

Besedna vrsta	≥ 3	≥ 2	≥ 1
Samostalnik	6.010	8.273	16.987
Pridevnik	2.873	3.794	7.466
Glagol	1.943	2.567	5.079
Vse	11.633	15.528	30.828

Tabela 2: Število različnih lem v korpusu DSI 2003−2006

3.1. Dostop do korpusa

Za uporabo korpusa potrebujemo programska orodja, predvsem konkordančnik. Dobri konkordančniki omogočajo iskanje po kombinacijah različnih kriterijev in znajo rezultate poizvedb prikazati na več načinov. Najbolj udobni in za resno delo najbolj primerni so konkordančniki, ki si jih instaliramo na lasten računalnik, vanje uvozimo korpus in ga analiziramo. Pri našem delu smo testno uporabili orodje Wordsmith (Scott, 2006), ki sicer ponuja poleg izdelave konkordanc tudi frekvenčne sezname, sezname ključnih besed in kolokacij, vendar pa deluje samo nad izvornim besedilom, kar pomeni, da ne moremo iskati po lemah ali oblikoslovnih oznakah, niti po oznakah zahtevati izpisa. Druga "slabost" orodja je, da je na voljo samo proti plačilu (obstaja pa tudi 40-dnevna evaluacijska verzija), korpus pa bo uporaben samo tistim, ki program kupijo in instalirajo.

Mrežni konkordančniki sicer nudijo manj možnosti, so pa dostopni vsem in uporabni brez posebne lokalne programske opreme. Na IJS že vrsto let obstaja konkordančnik na naslovu http://nl2.ijs.si/. Tu je na voljo več dvo- in enojezičnih korpusov slovenskega jezika, tudi (vsako leto obnavljani) DSI. Spletni vmesnik omogoča iskanje po korpusu in lahko prikaže rezultate na tri načine: kot spisek konkordanc, kot seznam zadetkov s frekvencami in, za dvojezične korpuse, kot seznam poravnanih segmentov. Primer rezultatov poizvedbe za pridevniki, ki jim sledi lema »aplikacija«, ki je izpisan kot frekvenčni seznam, je podan v Sliki 1.

Spletni vmesnik za iskanje uporablja strežnik korpusov CQP (Christ 1998), ki ima zelo bogat iskalni jezik, saj lahko prek regularnih izrazov (npr. »*aplikacij.**«) iščemo po kombinaciji pojavnic iz besedila, ali pa po njihovih oznakah (v našem primeru leme in oblikoslovne oznake MULTEXT-East). Ker je računalnik, na katerem teče servis, razmeroma močan, CQP pa optimiran na hitrost, tudi kompleksne poizvedbe vrnejo rezultat v razmeroma kratkem času.

Slabosti trenutne implementacije, ki se jih sicer zavedamo, bi pa v njihovo odpravljanje bilo treba vložiti nekaj dela, je zahteven jezik poizvedb,[1] ki bi ga bilo bolje prevesti v bolj strukturiran obrazec HTML, ter majhna možnost izbire oblike izpisa.

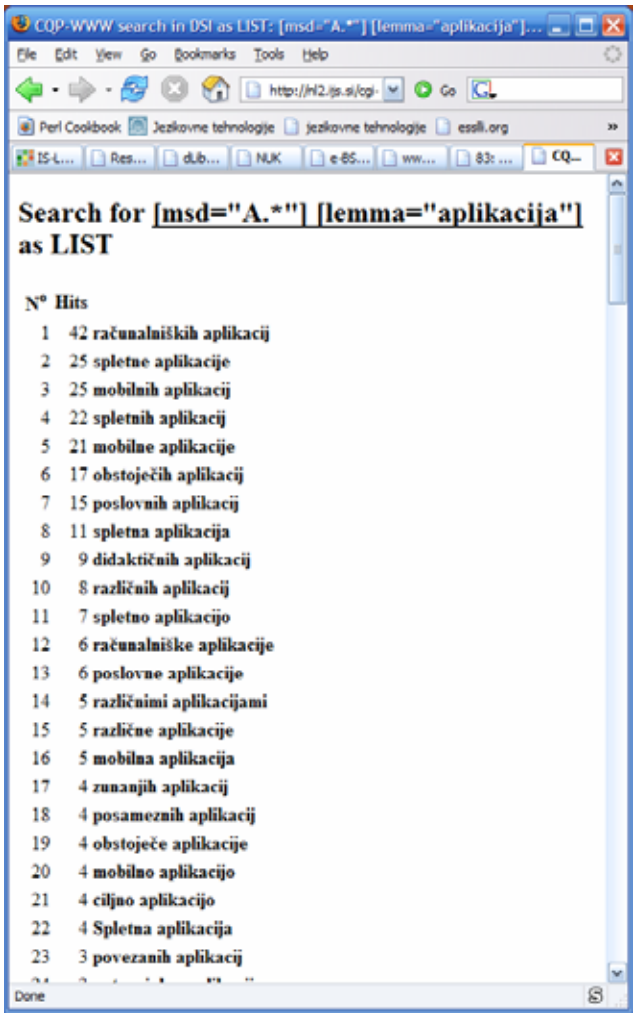
4. Uporaba korpusa DSI

Korpus DSI lahko uporabljamo za številne analize, ki nam pomagajo pri urejanju Islovarja. Pri pregledovanju primerov uporabe, ki so razvidni iz kolokacij, se odločamo o izboru slovenskih ustreznic in vrednotenju sinonimov, podatki o pogostosti posameznih izrazov pa nam pomagajo pri odločanju, katere izraze urediti prioritetno in katere sploh uvrstiti v Islovar.

4.1. Izbor slovenskih ustreznic

Prikazali bomo dva primera, kjer smo se pri odločanju o zapisu v Islovar odločali na podlagi korpusa DSI. Gre za izraze, ki se v informatiki pogosto uporabljajo, zlasti v besednih zvezah, pojavljajo pa se v več sinonimih.

[1] Poizvedba, ki poišče pridevnike v primerniku in besedne oblike od leme, ki se začne na *man* oz. *men* je *[msd="A.c.*"] [lemma="m[ae]n.*"]*



Slika 1. Primer rezultatov poizvedbe v korpusu DSI

Ker je za področje informatike osnovni jezik angleščina, se med domačim izrazjem pojavljajo tudi besede, privzete iz angleščine in v angleški pisavi. Slovenščina je sorazmerno negostoljubna do takih pojavov. Vprašanje je, ali take izraze zajeti v terminološki slovar in kako.

4.1.1. Online v slovenščini

Online se uporablja v angleškem jeziku kot pridevnik ali kot prislov, v inačicah *on-line*, *on line*, *online*. Ima širok pomen. Beseda v angleški pisavi obstaja v številnih slovenskih besedilih, tudi uglednih institucij, kot je COBISS (online informacijski sistem in servisi), Centralna tehniška knjižnica (online informacijski servis). Angleško-slovenski slovar bibliotekarske terminologije pa besedo zapiše *onlajn* (onlajn servis).

V elektronskih virih se online, on-line pojavljata v slovenskih besedilih z naslednjo pogostostjo:

Nova beseda	*najdi.si*	*korpus DSI*
352	517.000	102

Tabela 3: Pogostost online, on-line v elektronskih virih

Primeri iz korpusa Nova beseda niso uporabni, zvečine gre za imena časopisov, institucij. Iz Monitorja izvirata *on-line* (6) in *on line* (159), vendar gre za zapise iz let 2000 in 2002. Taki viri so za danes tako aktualen izraz zastareli.

Primeri uporabe v najdi.si so številni in zelo različni. Sicer pa je pregled takšne množice, tudi če jo skrčimo po raznih kriterijih, praktično nemogoč. Iz teh podatkov lahko sklepamo samo to, da je beseda *online* v slovenščini zelo pogosta, po vsej verjetnosti zato, ker ji nismo našli prave ustreznice.

Korpus DSI nam omogoča boljšo preglednost uporabe in navaja pojavnosti, kot so:

- online učenje na daljavo, online literarna revija, online storitve, online katalog, online prijava;
- on-line prodaja, on-line sestanek, on-line nakup, on-line poslovanje.

Očitno je pomen angleške besede *online* tako širok, da ga je slovenščini nemogoče nadomestiti z eno samo ustreznico. Izraz se je iz strokovne rabe prenesel tudi v splošno besedišče. Sorazmerno redko se nadomešča z drugimi slovenskimi ustreznicami. Torej smo se odločili, da ga vključimo v Islovar v angleški pisavi, urejenega v naslednji slovarski sestavek:

online neskl. *(angl. online, on-line)* žarg.
1. ki je dostopen v oddaljenem računalniškem sistemu, npr. online podatkovna baza
2. ki je dostopen po telekomunikacijskem omrežju, npr. online banka
3. gl. povezan in priključen in priklopljen
4. gl. spleten
5. gl. sproten

V Islovarju imamo zdaj urejenih 25 slovarskih sestavkov, ki se nanašajo na angleški *online*. Ti sestavki se bodo z leti brez dvoma pomnožili in tudi spremenili, ko bo slovenščina ta izraz povsem absorbirala ali pa ga izločila. Za take spremembe je Islovar odlično opremljen, saj omogoča takojšnje posodabljanje.

4.1.2. Management v slovenščini

Management je beseda, ki se je uveljavila v slovenščini v zadnjih petnajstih letih v pomenih: poslovodstvo, upravljanje, vodenje. Slovar slovenskega knjižnega jezika (SSKJ) pozna besedo še v angleški pisavi, Slovenski pravopis iz leta 2000 pa samo še kot *menedžment*, kar usmerja na vodenje, upravljanje.

V informatiki se ta izraz pojavlja v številnih besednih zvezah, v angleški pisavi pa tudi kot razne slovenske ustreznice, npr. upravljanje, obvladovanje, ravnanje. Za isti pojem se pojavlja več inačic, npr.: avtorji slovenijo *business process management* kot upravljanje poslovnih procesov, management poslovnih procesov, obvladovanje poslovnih procesov, sistem za procesno ravnanje.

V Islovarju smo imeli pred končnim urejanjem zbranih 53 iztočnic z ustreznicami za *management*, od teh številne sinonime ali po našem mnenju nepravilne ustreznice. Pri urejanju slovarskih sestavkov smo se naslonili na opredelitev splošnih pomenov v SSKJ in na primere uporabe in na pogostost v korpusu DSI.

Če s funkcijo »wordlist« izpišemo število vseh pojavitev v korpusu DSI, se nam potrdi domneva, da je *upravljanje* v informatiki povsem uveljavljen izraz. Zato smo ga razen v nekaterih primerih, zlasti pri izrazu *obvladovanje*, prevzeli tudi v Islovarju kot nadrejeni sinonim. *Menedžment* se v korpusu DSI pojavlja pretežno v pomenu poslovodstvo.

Izraz	*pogostost*
upravljanje	1207
obvladovanje	314
management	203
ravnanje	55
menedžment	8

Tabela 4: pogostost ustreznic za *management* v korpusu DSI

Pri zvezah z *upravljanje* smo opazili pogosto (346 krat) napačno rabo z orodnikom, npr. upravljanje z vsebinami, s tveganji, z znanjem, kar po SSKJ in Slovenskem pravopisu ni pravilno. V Islovarju te rabe zdaj sicer ne priporočamo, se bo pa morda v informatiki uveljavila.

V Islovarju imamo torej urejena naslednja slovarska sestavka:

management -a m *(angl. management)*
1. gl. menedžment (1) in upravljanje (3)
2. gl. vodenje
3. gl. poslovodstvo in uprava

upravljanje -a s *(angl. management)*
1. načrtovanje, nadziranje in vzdrževanje informacijske tehnologije, npr. upravljanje podatkovnih baz, upravljanje dokumentov
2. usmerjanje procesov, postopkov v organizaciji z uporabo informacijske tehnologije, npr. upravljanje znanja; sin. ravnanje (3)
3. organizacijska funkcija, katere osrednje naloge so načrtovanje, organiziranje, nadziranje dejavnosti; sin. management (1), menedžment (1)
4. računalniško usmerjanje delovanja sistema, naprave; prim. krmiljenje.

Beseda *upravljanje* je iz splošnega jezika prešla v strokovni jezik informatike, kjer se je uveljavila bolj kot angleška beseda *management* ali *menedžment.* Privzela je številne pomene. *Management* in *menedžment* pa se uporabljata v splošnem jeziku najpogosteje v pomenu poslovodstvo, uprava.

4.2. Pogostost izrazov kot osnova za urejanje

Analiza pogostosti informacijskih izrazov je zelo koristna, zlasti pri pregledu samostalniških besed. Kot smo že omenili, smo pri analizi korpusa DSI uporabili tudi orodje Wordsmith (Scott, 2006), ki poleg izdelave konkordanc in frekvenčnega seznama ponuja tudi izdelavo seznama ključnih besed. Te dobimo tako, da izberemo frekvenčna seznama našega (specializiranega) korpusa in korpusa splošnega jezika (v našem primeru je bil to vzorec iz korpusa FIDA), nato pa Wordsmith prek statističnih mer določi, katere besede (in s kakšno mero »ključnosti«) se v specializiranem korpusu pojavijo večkrat kot pričakovano glede na splošno besedišče.

Začetek seznam ključnih besedah korpusa DSI, urejenega po ključnosti, je podan v Sliki 2. V prvem stolpcu je ključna beseda, v drugem absolutna frekvenca v korpusu DSI, v četrtem v vzorcu FIDA, v petem pa vrednost ključnosti.

V seznamu sicer vidimo, da se v korpusu večkrat uporabljajo nekatere funkcijske besede, verjetno zaradi drugačnosti stila člankov od pretežno neznanstvenih besedil v FIDI, ter dosti več angleškega izrazja, za nas bolj zanimivi pa so samostalniki, kjer je seznam lepo razporejen. Izrazi, ki so na vrhu, so zares temeljni za informatiko. Bi pa seznam seveda bil bistveno bolj uporaben, če bi v njem lahko opazovali leme namesto besednih oblik.

Ključna beseda	Frekv. DSI	Frekv. Ref.	Ključnost
podatkov	2954	461	3238,9
sistema	1941	154	2652,98
procesov	1426	22	2437,37
storitev	1598	89	2354,81
sistem	1782	212	2165,91
poslovnih	1377	55	2141,23
podjetja	1757	331	1764,5
it	1019	22	1697,38
of	1277	118	1677,04
potrebno	1547	292	1551,94
informacijske	878	8	1543,88
poslovanja	983	48	1481,91
rešitev	1285	182	1464,9
and	959	60	1381,42
uporabnikov	816	20	1343,46
sistemov	897	48	1331,06
omogoča	1142	153	1329,91
rešitve	978	86	1301,59
informacij	1031	111	1294,21
informacijskih	713	3	1285,4
upravljanje	827	39	1253,79
uporabo	1014	118	1241,22
procesa	790	34	1214,86
projekta	992	117	1208,82
programske	730	26	1152,6
is	752	35	1142,46
tehnologije	705	27	1102,41
opreme	781	57	1088,24
ter	2924	1676	1075,16
uporabe	791	66	1067,11
ikt	570	0	1056,18
uporabniki	670	27	1040,24
informacijski	586	15	960,614
informatike	551	7	952,813
informacijskega	537	5	943,249
npr	813	111	939,883
spletnih	556	11	932,802
znanja	748	84	926,303

Slika 2. Ključne besede v korpusu DSI

Izrazi, ki se najpogosteje uporabljajo v pisani obliki, bi morali biti v Islovarju prioritetno urejeni. Pri pregledu opažamo, da so nekateri sicer urejeni, niso pa urejene vse besedne zveze, ki so že v Islovarju, verjetno bo treba še nekatere dodati. Prav te lahko najdemo iz primerov uporabe, ki jih najdemo s konkordančnikom v korpusu DSI.

V primerjavi s korpusom DSI iz leta 2005 se je rang nekaterih izrazov spremenil, kar dokazuje dinamičnost njihove uporabe. Upravljanje se je pojavilo 911 krat v letu 2005, v letu 2006 pa kar 1316 krat. Povečala se je tudi uporaba besede *management* s 303 krat na 450 krat. Zelo veliko se uporablja *aplikacija* (ki je Islovar ne priporoča, temveč usmerja na uporabniški program): v korpusu 2003−2005 729 krat, korpusu 2003−2006 pa kar 1412 krat. Na temelju teh ugotovitev bo verjetno treba popraviti slovarski sestavek za *aplikacijo*.

Pridevniške besede nastopajo v terminološkem slovarju predvsem v besednih zvezah. Iz korpusa DSI smo izluščili 142 tipičnih najpogostejših pridevniških besed, pri katerih opažamo podobno razporeditev kot pri samostalnikih. Na vrhu so: *informacijski* (3562 krat), *spletni* (1752 krat), *elektronski* (1670 krat), *podatkovni* (1200 krat). S temi pridevniškimi besedami lahko poiščemo vse možne besedne zveze, ki nastopajo v korpusu in jih primerjamo z vsebino Islovarja.

V Islovarju že najdemo 28 različnih zvez s *podatkovni*, npr. *podatkovna baza*, *podatkovno skladišče*, *podatkovno rudarjenje*. V korpusu DSI pa so še številne druge (skupno 142), npr. *podatkovni agregat, podatkovni element, podatkovni strežnik, podatkovno upravljanje*. Z analizo uporabe presodimo, katere od teh še vključiti v Islovar. Na ta način dopolnjujemo vsebino Islovarja z aktualnimi izrazi.

5. Uporaba drugih elektronskih virov

Zborniki DSI zajemajo predvsem organizacijsko informacijske teme, ki so v času odvijanja posvetovanja aktualne. Zato korpusa DSI ne moremo uporabiti za odločanje o številnem izrazju, ki tudi sodi v Islovar.

Navedimo kot primer izraz *pomnilnik*, ki se zaradi razvoja informacijske tehnologije pojavlja v številnih različicah in zvezah, v Islovarju kot iztočnica ali del besedne zveze kar 111 krat, v razlagah pa 146 krat. V korpusu DSI ga najdemo samo 37 krat v nekaterih splošnih pomenih. Tukaj se moramo usmeriti na druge elektronske vire, predvsem na iskalnika najdi.si in Google. Pregledovanje je v tem primeru zelo zamudno, ker oba iskalnika navajata veliko število primerov, med katerimi so pretežno reklamna besedila prodajalcev.

Korpus Nova beseda navaja pomnilnik 2619 krat, primeri pa so iz časnika Delo in revije Monitor. Delo navaja *pomnilnik* v najbolj splošnem pomenu. Monitor je zanesljiv vir, vendar v tem primeru slabo uporaben, ker so primeri iz leta 2000 – torej očitno zastareli.

Nova zbirka Besede slovenskega jezika najde primere za *pomnilnik* 13 krat, pretežno iz najdi.si. Med drugim navaja *brisljiv*, *nebrisljiv pomnilnik* iz istega, vendar nezanesljivega vira.

Monitor izdaja zdaj tudi skrajšano, spletno inačico revije, ki pa žal še nima velikega arhiva. Podobno revija Moj mikro. Zato smo morali pri zbiranju izrazov uporabiti številne, tudi knjižne vire. Ker pa se tehnologija razvija, se bo družina *pomnilnik* v Islovarju brez dvoma še razširjala.

Ugotavljamo, da sta Google in najdi.si dobri orodji predvsem za ugotavljanje pojavnosti nekega izraza v spletnih dokumentih. Če ugotovimo, da se neki izraz pojavlja večkrat v zanesljivih virih, potem lahko presodimo, da ta izraz sodi v Islovar kot iztočnica, kot enakovreden ali nadrejeni sinonim.

6. Sklepne ugotovitve

Pri zbiranju in urejanju izrazja za terminološki slovar informatike se uredniki naslanjajo predvsem na dostopne elektronske informacijske vire. Pomembna kriterija pri odločanju o uporabi teh virov sta ažurnost in zanesljivost.

Pri iskanju pogostosti uporabe posameznih izrazov in opredeljevanju pomena pri oblikovanju razlage si uredniki pomagajo s slovarji in dokumenti, ki so dostopni na svetovnem spletu. Pri teh raziskavah je v veliko pomoč tudi specializirani korpus DSI, ki omogoča različne vpoglede v uporabo in se posodablja vsako leto.

Kot smo pokazali na dveh primerih urejanja, nam uporaba tega korpusa zanesljivo pokaže pogostost in pomene posameznih izrazov v informatiki v sedanjem času. Namen spletnega terminološkega slovarja Islovar pa je med drugim tudi zasledovati in beležiti razvoj informacijskega izrazja v slovenščini.

Korpus DSI za zdaj zajema zbornike posvetovanj Dnevi slovenske informatike. Lahko bi ga razširili z različnimi besedili, npr. s članki strokovnih revij in z besedili učbenikov ter drugih publikacij, tako da bi postal bolj uravnotežen in še bolj uporaben. To pa bi bil zahtevnejši projekt, ki bi potreboval širšo podporo.

Literatura in viri

Arhar, Š.(2006): Gradnja specializiranega korpusa. Jezik in slovstvo, 2006, št. 1

Christ, O. (1994): A modular and flexible architecture for an integrated corpus query system. Proceedings of the 3rd Conference of Computational Lexicography and Text Research (COMPLEX'94), Budimpešta. *http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/*

Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. (2005). Massive multilingual corpus compilation: ACQUIS Communautaire and totale. *Proceedings of the Second Language Technology Conference*, april 2004, Poznan.

Erjavec, T., Vintar, Š. (2004). Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika. *Uporabna informatika. (Ljubljana)*, 12/2 97-106.

Pahor, D., et al.(2002): Leksikon računalništva in informatike, Ljubljana, Založba Pasadena

Scott, M. (2006): WordSmith Tools, Version 4. Oxford University Press. *http://www.lexically.net/wordsmith/*

Turk, T., Puc, K. (2006): Islovar kot model spletnega terminološkega slovarja. Razvoj slovenskega knjižnega jezika (Obdobja 24 – Metode in zvrsti) (v tisku)

Slovarski knjižni viri:

Slovenski pravopis, Ljubljana: Založba ZRC, ZRC SAZU, 2001

Angleško-slovenski slovar bibliotekarske terminologije, Ljubljana: Narodna in univerzitetna knjižnica, 2002

Spletni viri:

Islovar *http://islovar.org*

Slovar slovenskega knjižnega jezika *http://bos.zrc-sazu.si/sskj.html*

Besede slovenskega jezika *http://bos.zrc-sazu.si/besede.html*

Nova beseda *http://bos.zrc-sazu.si/s_beseda.html*

Spletni slovarji *http://www.sigov.si/slovar.html*

Zbirka tujih slovarjev One Look *http://onelook.com*

Konkordančnik za korpus DSI: A WWW Concordance Service *http://nl2.ijs.si/index-mono.html*

Evroterm *http://www.gov.si/evroterm/*

Spletne strani:

Najdi.si *http://www.najdi.si*

Google *http://google.com/*

Monitor *http://www.monitor.si/*

Moj mikro *http://www.mojmikro.si/index.plus*