Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika

Tomaž Erjavec
Odsek za tehnologije znanja, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
Špela Vintar
Oddelek za prevajalstvo, Filozofska fakulteta, Aškerčeva 2, 1000 Ljubljana
spela.vintar@guest.arnes.si

Povzetek

Prispevek predstavi uporabo zbirke besedil (jezikovnega korpusa) pri izdelavi terminološkega slovarja. Spletni slovar informacijskega izrazja slovenskega jezika nastaja pri jezikovni sekciji Slovenskega društva informatika (SDI), društvo pa organizira tudi letne konference »Dnevi slovenske informatike« (DSI) s tiskanimi zborniki. V prispevku najprej predstavimo slovar, nato pa se osredotočimo na izgradnjo korpusa s področja informatike, ki trenutno zajema zbornik konference DSI 2003. Izdelava korpusa temelji na uporabi tehnologij XML in je sestavljena iz pretvorbe prispevkov v zborniku iz izvornega zapisa (Microsoft Word) v osnovni zapis XML, nato pa v obliko, primerno za spletno iskanje. Manjši del korpusa je dvojezični in vsebuje slovenske in angleške stavčno poravnane povzetke prispevkov. Izvorni namen izdelave korpusa DSI je slovaropisni, saj bi z njim po eni strani želeli sodelujočim olajšati izdelavo slovarja SDI, po drugi strani pa ponuditi uporabnikom dodatni vir primerov za iskani termin. V članku opišemo postopke izdelave korpusa in računalniško podprtega iskanja izrazov, pri katerem so sodelovali tudi študentje prevajalstva na Filozofski fakulteti Univerze v Ljubljani. Prispevek obravnava tudi načrte za nadaljnje delo, ki poleg razširitve korpusa predvidevajo tudi oblikoskladenjsko označevanje in lematizacijo besed v korpusu ter avtomatsko luščenje področnih terminov.

Abstract

A Corpus-driven Approach to Building the dictionary of Information Science

The paper describes the exploitation of a text corpus for the compilation of the terminological dictionary of information science, which is being created within the language section of the Slovenian Society of Information Science (SDI). Among its other activities, the Society organizes yearly meetings under the title »Days of Slovenian Information Science« (DSI) with printed proceedings. The first part of the paper presents the web dictionary and the process of building the corpus of information science, which at present contains the proceedings of the conference DSI 2003. Building the corpus included several stages, such as conversion of the original Word files into XML and transformation into a web-searcheable format. A small section of the corpus is bilingual and consists of English and Slovene sentence-aligned abstracts. The second part of the paper describes methods of corpus-based terminography, which were employed within a student project at the Faculty of Arts, Department of Translation. Finally, plans for future work, including deeper linguistic tagging, term extraction and corpus expansion are discussed.

1 UVOD

Korpus je zbirka besedil, ki so izbrana tako, da služijo kot vzorec za stanje, raznovrstnost ali razvoj nekega jezika. Uporaben je kot podlaga, na kateri gradimo opise jezika, ali pa kot sredstvo za preverjanje hipotez o jeziku. Čeprav so korpusi dandanes koristni pri številnih dejavnostih (jezikoslovje, poučevanje jezika, razvoj jezikovnih tehnologij), so se najprej uporabljali pri slovaropisju, kjer služijo kot vir primerov uporabe besed in besednih zvez. Tradicionalno so bili korpusi hranjeni na papirju (tipično v obliki listkov, od katerih je vsak navajal primere uporabe ene slovarske iztočnice), bistven premik na tem področju pa je naredil angleški slovar Cobuild (opisan v Sinclair 1998), saj je bil prvi slovar, ki je nastal izključno na podlagi računalniško hranjenega referenčnega korpusa The Birmingham Collection of Texts, iz katerega je pozneje nastala zbirka Bank of English (http://titania.cobuild.collins.co.uk/; v času pisanja ta korpus ni več dostopen za spletno iskanje. Računalniški korpus ima v primerjavi s klasičnim »papirnim« vrsto prednosti: hraniti je mogoče bistveno večjo količino besedil, ta je mogoče preprosto in hitro preiskovati po različnih kriterijih, rezultate poizvedb pa predstaviti prilagojeno specifičnemu namenu.

Od časov Cobuilda je uporaba računalniških korpusov pri izdelavi slovarjev postala že standardna praksa, ki se je z angleškega razširila tudi na jezike z manjšim številom govorcev. Za slovenski jezik je tak primer korpus FIDA (http://www.fida.net/), ki ga pri DZS, d. d., uporabljajo za izdelavo nove generacije slovarjev, na Filozofski fakulteti za jezikoslovne raziskave, na IJS in pri podjetju Amebis, d. o. o., pa za razvoj jezikovnih tehnologij. Bank of English in FIDA spadata med t. i. referenčne korpuse, katerih cilj je čim bolje vzorčiti celotno produkcijo nekega jezika. Referenčni korpusi so tipično zelo veliki (FIDA ima sto milijonov besed) in vsebujejo veliko število besedil (FIDA prek 20.000), ki so izbrana po skrbno uravnovešeni mreži kriterijev, s katero je na primer določeno razmerje med leposlovjem in strokovno literaturo, monografijami in periodiko, izvorno in prevodno literaturo itd. Poleg referenčnih korpusov pa poznamo tudi specializirane korpuse, ki so usmerjeni samo v določen segment jezika oz. njegove uporabe, npr. jezik najstnikov (COLT) ali pa jezik poizvedb po letalskih poletih. Takšen korpus, tj. specializirani korpus informatike slovenskega jezika, bo tudi predmet tega članka.

Uporabnost nekega korpusa je odvisna od njegove velikosti pa tudi urejenosti, tj. kako podrobno je dokumentiran in označen, ter standardiziranosti njegovega zapisa. Dokumentiranost omogoča vpogled v vire, ki so bili uporabljeni za izgradnjo korpusa, kakšni uredniški posegi so bili narejeni nad temi viri, oznake, ki so uporabljene v korpusu itd. Označenost korpusa, npr. oblikoskladenjska na ravni stavka, besede, z oblikoskladenjskimi oznakami itd., omogoča bogatejše izkoriščanje korpusnega materiala, saj lahko po njem iščemo po bolj abstraktnih kategorijah, npr. »najdi vse pojavitve leme "aplikacija", pred katero stoji pridevnik«. Standardiziranost zapisa pa doprinese k izmenljivosti korpusa, tako med ljudmi kot med aplikacijami, in k neodvisnosti od konkretnih računalniških platform, s tem pa tudi k večji trajnosti. Standardiziranost dandanes pomeni v prvi vrsti zapis v skladu z XML – eXtended Markup Language (W3C 2000), saj je to edini ustrezni standard za zapis digitalnih besedil, ki je tudi široko podprt v programski opremi in pridruženih standardih. V nadaljevanju se vrnemo k tem temam in opišemo naše rešitve pri izgradnji korpusa.

V prispevku predstavimo jezikovni korpus, ki služi kot podpora pri izdelavi spletnega slovarja informacijskega izrazja slovenskega jezika, nastajajočega pri jezikovni sekciji Slovenskega društva Informatika (SDI), ter uporabljene korpusno-terminološke metode za pridobivanje izrazja. V drugem poglavju na kratko predstavimo slovar, tretje poglavje opiše izdelavo našega korpusa, njegov zapis, možnosti nadaljnjega označevanja ter mrežni konkordančnik, s katerim lahko iščemo po korpusu, četrto poglavje opiše terminološko delo, ki smo ga v preskusne in izobraževalne namene izvajali s študenti prevajalstva, peto poglavje pa poda nekaj sklepov in načrte za nadaljnje delo.

2 SLOVAR SLOVENSKEGA DRUŠTVA INFORMATIKA

Slovensko društvo Informatika je v okviru svoje jezikovne sekcije leta 2001 začelo z delom na spletnem slovarju informacijskega izrazja slovenskega jezika, na kratko »Slovar informatike«. Slovar, ki se nahaja na naslovu http://www.ef.uni-lj.si/terminoloskislovar/, je namenjen vsem članom društva in široki javnosti. V slovarju se zbirajo temeljni in najsodobnejši informacijski izrazi, ki se uporabljajo v znanosti, v strokovni javnosti in med uporabniki. Pomagal naj bi pojasnjevati pomen strokovnih pojmov vsem, ki se srečujejo z informatiko, pa tudi pri ustvarjanju znanstvenih del, pri pisanju strokovnih besedil in pri komuniciranju z uporabniki.

Slovar se sproti dopolnjuje neposredno na spletu. Pri njegovem oblikovanju sodelujejo številni strokovnjaki kot uredniki področij, strokovni sodelavci, svetovalci ali kot člani sekcije. Za zdaj ima opredeljenih 16 področij informatike, npr. internet, poslovna informatika, varovanje informacijskih sistemov, naprave (strojna oprema), sociološki vidiki, odprti sistemi itd. Terminološki slovar informatike je razlagalni in informativni slovar, ki strokovno izrazje pomensko in jezikovno opisuje, vrednoti in kateremu so dodani angleški ustrezniki. Slovarski sestavek oblikujejo iztočnica (enobesedni ali večbesedni izraz), besednovrstna in stilska oznaka, ustreznik v angleškem jeziku in razlaga, ki jo lahko podkrepijo sinonimi in vsebinsko povezani pojmi, ki so obravnavani v slovarju. Za dokončno vsebino in oblikovanje slovarja so zadolženi uredniki. Ti izraze in razlage preverjajo glede na že obstoječe, objavljeno izrazje, pa tudi v skladu s pravili slovenskega jezika.

Uporabniki slovarja lahko iščejo slovenske ali angleške besede, lahko pa nove slovenske besede ali prevode tudi vpisujejo. Slovar se naslanja na angleške izraze, zato vpis slovenskega izraza brez angleškega

ni mogoči. Vsi izrazi, ki jih uporabniki ne najdejo, se beležijo in po presoji uredništva vnašajo v slovar. Slovar zajema samo informacijsko izrazje, besed splošnega pomena ne vsebuje.

3 KORPUS DNEVOV SLOVENSKE INFORMATIKE

Slovensko društvo Informatika organizira letne konference »Dnevi slovenske informatike« (DSI) s tiskanimi zborniki. Ker zborniki pokrivajo isto področje kot slovar, obenem pa so znanstveni prispevki dragocen vir svežega slovenskega izrazja, se je pojavila ideja, da se zbornike pretvori v korpus, ki bi nato lahko služil kot podpora pri izdelavi, pa tudi uporabi slovarja.

Vir za izdelavo korpusa, ki ga predstavimo v tem razdelku, so digitalni izvirniki posameznih prispevkov (torej brez predgovorov in drugega spremnega besedila v zborniku), ki so služili kot predloga tiskanemu zborniku za leto 2003. Pri izdelavi korpusa smo izhajali iz določenih standardov; tako za zapis korpusa uporabljamo XML (W3C 2000), za pretvorbe pa pridruženi standard XSLT (W3C 1999). Če bo korpus v prihodnosti prerasel svojo sedanjo namembnost in velikost, načrtujemo tudi prilagoditev korpusa priporočilom Iniciative za zapis besedil TEI – Guidelines for Text Encoding and Interchange (Sperberg-McQueen and Burnard, 2002).

3.1 Opis vira

Triindevetdeset člankov, ki so služili kot osnova korpusu, je zapisanih v formatu Microsoft Word, pri čemer je stil predpisan s strani SDI. Stil je sicer podan opisno, spletne strani z navodili za avtorje pa ponujajo tudi primer pravilno oblikovanega članka. Predloga ponuja poleg standardnih tudi svoje stile, ki definirajo nekatere strukturno pomembne dele članka, kot so npr. naslov, avtorji, njihovi naslovi, slovenski in angleški povzetek itd. Uporaba takšnih stilov zelo olajša pretvorbo v korpus, žal pa jih avtorji niso upoštevali, čeprav je v urejevalniku word mogoče doseči isto podobo besedila z različnimi prijemi.

3.2 Pretvorba v XML

Za pretvorbo oblike Microsoft Word v XML obstaja razmeroma bogata ponudba večinoma komericalnih programov. Mi smo izbrali program UpCast (http://www.infinity-loop.de/), ki v prosto dostopni »osebni licenci« ponuja polno funkcionalnost pri pretvorbi dokumentov, je pa potrebno za vsak dokument posebej sprožiti pretvorbo. To delo smo zaupali študentom prevajalstva, ki so v okviru predmeta korpusi in saj podatkov v tretjem letniku dodiplomskega študija izdelovali korpuse za terminografske namene.

```
<article>
  <para role="naslov_prispevka">JEZIKOVNI VIRI SLOVENSKEGA
    STROKOVNEGA JEZIKA</para>
  <para role="avtor">Tomaž Erjavec</para>
  <para role="avtor_naslov">Odsek za inteligente sisteme, Institut "Jožef
    Stefan", Jamova 39, 1000 Ljubljana</para>
  <para role="avtor_naslov">tomaz.erjavec@ijs.si</para>
  <para role="povzetek_naslov">Povzetek</para>
  <para role="povzetek">Prispevek predstavi področje jezikovnih tehnologij,
    metod, ki olajšajo uporabo jezika v ...</para>
  <para role="abstract_title">Abstract</para>
  <para role="abstract">
    <phrase>LANGUAGE RESOURCES FOR SLOVENE TECHNICAL
      LANGUAGE</phrase>
  </para>
  <para role="abstract">The paper discusses the field of Language
    Technologies, i.e. methods that ...</para>
  <section>
    <title>
      <phrase role="upcast-HEADINGNUMBER">1.</phrase>
      UVOD
    </title>
    <para role="Normal">
      Prispevek predstavi po
      <phrase>dročje jezikovnih tehnologij: metod, ki ...
```

Slika 1: **Primer pretvorbe iz Worda v XML z orodjem upCast**

```
<s><w>Večina</w> <w>sorodnih</w> <w>člankov</w><c>,</c> <w>ki</w>
<w>smo</w> <w>jih</w> <w>zasledili</w><c>,</c> <w>obravnava</w> <w>le</w>
<w>algoritme</w> <w type="abbr">oz.</w> <w>postopke</w> <w>za</w>
<w>razvrščanje</w> <w>besedil</w> <w>kot</w> <w>v</w> <w>članku</w>
<c type="open">[</c><w type="dig">15</w><c type="close">]</c> <w>ali</w>
<c type="open">[</c><w type="dig">14</w><c type="close">]</c><c>.</c></s>
```

Slika 2: **Primer stavka iz korpusa DSI**

UpCast ponuja izhod v lastnem tipu dokumentov XML, eksperimentalno pa tudi v zapisu DocBook (http://www.docbook.org/), ki se sicer uporablja predvsem za zapis računalniške dokumentacije, je pa dovolj pregleden pa tudi dobro dokumentiran; zapis začetka enega od prispevkov v tem izhodnem formatu ilustriramo v sliki 1.

Kot vidimo, ima format precejšnje število koristnih podatkov, čeprav ni brez pomankljivosti, tako je npr. v naslovu nepojasnjeno ena od črk mala, v besedilu pa se brez prave logike pojavi element <phrase>. Kakorkoli že, s to pretvorbo preidemo v standardiziran in poenoten format (XML DocBook), ki je z uporabo primernega stila – vsaj teoretično – še vedno prikazljiv enako kot original in torej ni izgubil informacije.

3.3 Jezikovno označevanje

Po pretvorbi v enoten zapis TEI lahko zbirko besedil že poimenujemo korpus, saj je uniformno in standardizirano zapisan. Seveda pa se s tem prava jezikovna analiza besedila šele začne. Kaj točno hočemo v korpusu označiti, je v veliki meri odvisno od namembnosti. Osnovna koraka, ki sta vedno koristna, sta označitev besed in stavkov v besedilu, t. i. tokenizacija in segmentacija. Čeprav že ta stopnja označevanja skriva določene pasti (pika npr. ne označuje vedno konca stavka), pa v splošnem ni preveč zahtevna – to je tudi stopnja, do katere smo trenutno označili korpus DSI, primer stavka iz korpusa pa podamo v sliki 2.

Naslednja faza, ki je pogosto koristna, je t. i. oblikoskladenjsko označevanje (Van Halteren, 1999): tu vsaki besedi v korpusu pripišemo njene oblikoskladenjske oznake, npr. »samostalnik moškega spola v rodilniku ednine«, dostikrat pa tudi leme oz. gesla, npr. za besedo »berači« lemo »beračiti«. Za takšno označevanje je potrebno najprej imeti slovar ali pa program, ki za besedne oblike določi vse možne oblikoskladenjske oznake in po možnosti pripadajoče leme. Neka besedna oblika ima v slovarju ponavadi več možnih interpretacij, tako je npr. »berači« lahko glagol v velelniku ali povedniku ali samostalnik v imenovalniku ali orodniku množine. V konkretnem besedilu pa bo besedna oblika imela seveda samo eno ustrezno oznako. Naloga programov za oblikoskladenjsko označevanje je izmed možnih oblikoskladenjskih oznak neke besede določiti glede na sobesedilo, njeno pravo oznako.

Izdelanih je bilo že veliko označevalnikov, ki se lahko naučijo zakonitosti nekega jezika iz ročno označenih korpusov. Ena bolj odmevnih metod z uporabo t. i. skritih markovskih verig določi najbolj verjetno zaporedje oblikoskladenjskih oznak besed v nekem stavku glede na njihov lokalni kontekst. Za angleški jezik dosežejo takšni označevalniki ob uporabi zadosti velike učne množice približno 96-odstotno natančnost. Za slovanske jezike, ki imajo precej bogatejše oblikoslovje in s tem večje število možnih oznak, je ta natančnost manjša, predvsem pa odvisna od velikosti učnega korpusa. Pri lastnih poskusih (Džeroski et al., 2000) smo dosegli natančnost reda 92 %. Kot primer rezultata tokenizacije, segmentacije in oblikoskladenjskega označevanja podamo v sliki 3 stavek iz korpusa MULTEXT-East (Erjavec, 2004).

3.4 Stavčna poravnava

Kot je v navadi za večino strokovnih publikacij, morajo tudi prispevki srečanja DSI vsebovati povzetek v

```
<s id="Osl.1.2.2.1">
<w lemma="biti" ana="Vcps-sma">Bil</w>
<w lemma="biti" ana="Vcip3s--n">je</w>
<w lemma="jasen" ana="Afpmsnn">jasen</w>
<c>,</c>
<w lemma="mrzel" ana="Afpmsnn">mrzel</w>
<w lemma="aprilski" ana="Aopmsn">aprilski</w>
<w lemma="dan" ana="Ncmsn">dan</w>
<w lemma="in" ana="Ccs">in</w>
<w lemma="ura" ana="Ncfpn">ure</w>
<w lemma="biti" ana="Vcip3p--n">so</w>
<w lemma="biti" ana="Vmps-pfa">bile</w>
<w lemma="trinajst" ana="Mcnpnl">trinajst</w>
<c>.</c>
</s>
```

Slika 3: **Stavek iz korpusa MULTEXT-East**

slovenskem in angleškem jeziku. Iz teh povzetkov je torej mogoče oblikovati vzporedni korpus, ki je za terminografske namene tudi najbolj uporaben tip korpusa. Ker je bil v izvirnih dokumentih za povzetka uporabljen poseben slog, smo povzetke iz besedil izluščili avtomatsko s pomočjo ustreznih oznak XML.

Stavčna poravnava je postopek, pri katerem se vsaki stavčni enoti izvirnika priredi ustrezna enota v prevodu. Postopek je delno avtomatiziran in ga zna opraviti tako rekoč vsak prevajalski program, vendar je rezultate samodejne poravnave navadno treba ročno pregledati in popraviti. Poravnava je tako predstavljala eno od študentskih opravil, zanjo pa smo uporabili prevajalski program DejaVu proizvajalca Atril (http://www.atril.com).

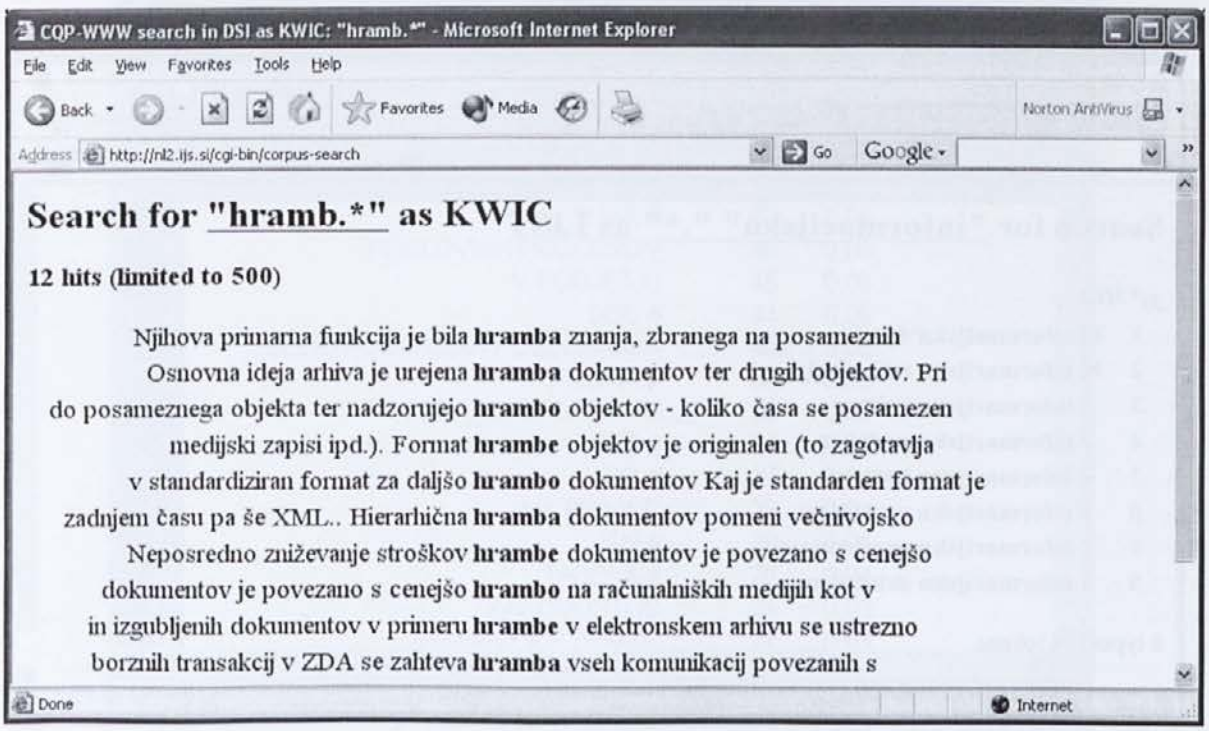
Postopek poravnave ustvari vzporedno besedilo, ki je na voljo bodisi v obliki dvostolpčne tabele v programu DejaVu, lahko pa ga izvozimo v MS Excel ali v besedilno datoteko, kjer sta izvirni in prevodni segment med seboj ločena s posebnim znakom, na primer s tabulatorjem.

3.5 Konkordance

Ko je korpus narejen, ga je seveda potrebno dati na razpolago. V našem primeru ciljno skupino uporabnikov, vsaj v prvi fazi, sestavljajo avtorji oz. uredniki slovarja, ki bi jim korpus pomagal pri preverjanju hipotez o slovenskih terminih. Za takšno delo se uporabljajo t. i konkordančniki, programi, ki prikažejo neko besedo ali besedno zvezo v vseh pojavitvah v korpusu skupaj s sobesedilom. Konkordančnik je temeljno orodje sodobnih slovaropiscev, saj ilustrira uporabo (in s tem posredno tudi pomene) iskanih besed ali besednih zvez. Poizvedovalni jeziki konkordančnikov so lahko precej bogati in obsegajo regularne izraze nad nizi, poizvedovanje glede na oznake ter logične operatorje.

V Sloveniji obstaja že večje število mrežnih konkordančnikov, na primer za referenčna korpusa FIDA (http://www.fida.net/) in Nova beseda (http://bos.zrc-sazu.si/) in za slovensko-angleški Evrokorpus (http://www.gov.si/evrokorpus/). Slika 4 prikaže izpis na poizvedbo v konkordančniku IJS; to orodje je dostopno na http://nl2.ijs.si/, ki ponuja večje število korpusov, sedaj tudi korpus DSI.

Konkordančnik IJS je v uporabi več kot štiri leta, uporabljajo pa ga predvsem prevajalci in študentje prevajanja. Kot hrbtenico uporablja IMS Corpus Workbench (CWB, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/), program za linux, ki je sposoben po kompleksnih kriterijih hitro iskati po



Slika 4: **Izpis enojezične konkordance**

velikih korpusih. CWB je nato prek skripta CGI postavljen na mrežo s HTTP strežnikom Apache. Poizvedovanje po izbranem korpusu poteka kar v iskalnem jeziku, ki ga ponuja CWB, ali pa v poenostavljenem načinu (npr. »info*«), ki se nato avtomatsko prevede v bolj kompleksno sintakso CWB (»"info.*"«). Izbrati je možno več načinov izpisa: poleg besede v kontekstu še vzporedni prikaz, primeren za dvojezične korpuse, in izpis golega seznama zadetkov, koristen npr. za iskanje sopojavnic – primer je podan v sliki 5.

Da mrežnemu konkordančniku dodamo nov korpus, kot smo to storili z DSI, je potrebno le-tega pretvoriti v vhodni format (kar je iz XML-ja z XSLT enostavno), ga tam indeksirati in dodati novo izbiro v katalog, skripto CGI ter krovno stran iskalnika.

Trenutno se korpus DSI prek mrežnega konkordančnika uporablja za iskanje novih terminov in preverjanja njihove uporabe s strani registriranih avtorjev slovarja. V bodoče nameravamo tudi dodati povezavo na neposredno iskanje terminov kot možnost za uporabnike slovarja.

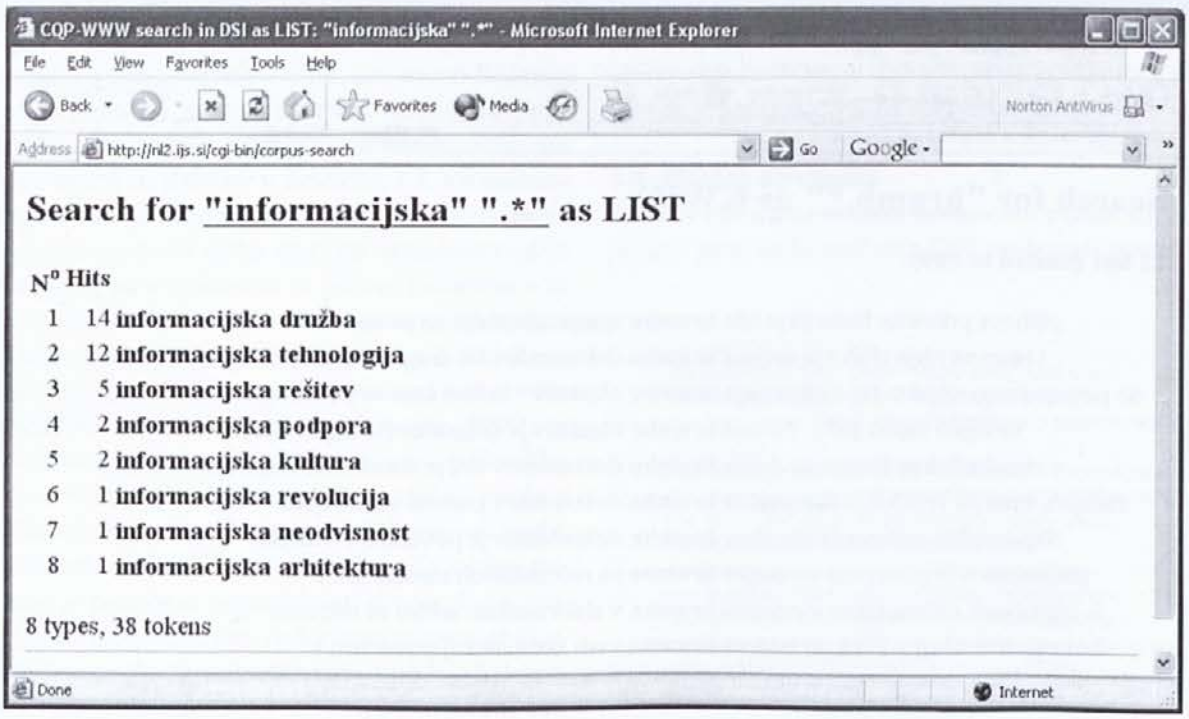
4 METODE KORPUSNO PODPRTE TERMINOGRAFIJE

V okviru študija prevajalstva že tretje leto poteka seminar korpusi in baze podatkov, pri katerem se študenti seznanijo z izdelavo korpusa za terminološke namene ter izdelujejo terminološke baze za različna področja. Jeseni 2003 je bila vzpostavljena naveza sodelovanja z društvom SDI, ki je pripeljala tudi do zamisli o sodelovanju študentov pri gradnji korpusa DSI in iSlovarja. Pred pričetkom dela je urednica slovarja Katarina Puc študentom predstavila značilnosti projekta, nato pa se je desetčlanska skupina študentov posvetila področju informatike.

Naloga je zajemala pretvorbo Wordovih datotek v XML z že opisanim orodjem UpCast, izdelavo dvojezičnega korpusa DSI s poravnavo, izbor gesel na podlagi obdelave s programom Wordsmith, slovarsko obdelavo gesel in nazadnje vnos obdelanih gesel v terminološki program TRADOS MultiTerm (http://www.trados.com). Ker smo prva dva koraka podrobno opisali že v prejšnjih razdelkih, se tu osredotočamo na postopke obdelave korpusa za terminološke namene.

4.1 Orodje Wordsmith

Čeprav je bil v času študentskega projekta že na voljo tudi mrežni konkordančnik, smo tu namesto njega uporabljali druga orodja, ki prvič omogočajo delo tudi brez internetne povezave, kar je za nekatere študente še vedno pomemben dejavnik, drugič pa poleg iskanja po



Slika 5: **Izpis v obliki seznama**

besedilih omogočajo tudi druge jezikovnotehnološke obdelave.

Pri ugotavljanju, katere besede ali besedne zveze so terminološko relevantne, si lahko pomagamo z orodjem Wordsmith (http://www.lexically.net), ki poleg brskanja po besedilni zbirki in izpisa konkordanc nudi še številne druge funkcije, na primer izdelavo besednih seznamov po pogostosti ali abecedi, in sicer posameznih besed ali večbesednih skupkov. Pri izdelavi seznamov je mogoče samodejno izločiti t. i. prazne besede, kot so vezniki, pomožni glagoli in podobno, s komponento Keywords pa lahko primerjamo dobljeni besedni seznam z referenčnim korpusom in ugotovimo, katere besede so s svojo relativno pogostostjo tipične za obravnavano stroko. Kot vsak boljši konkordančnik zna tudi Wordsmith izračunati kolokacije, pa tudi grafično prikazati distribucijo posameznega izraza v korpusu.

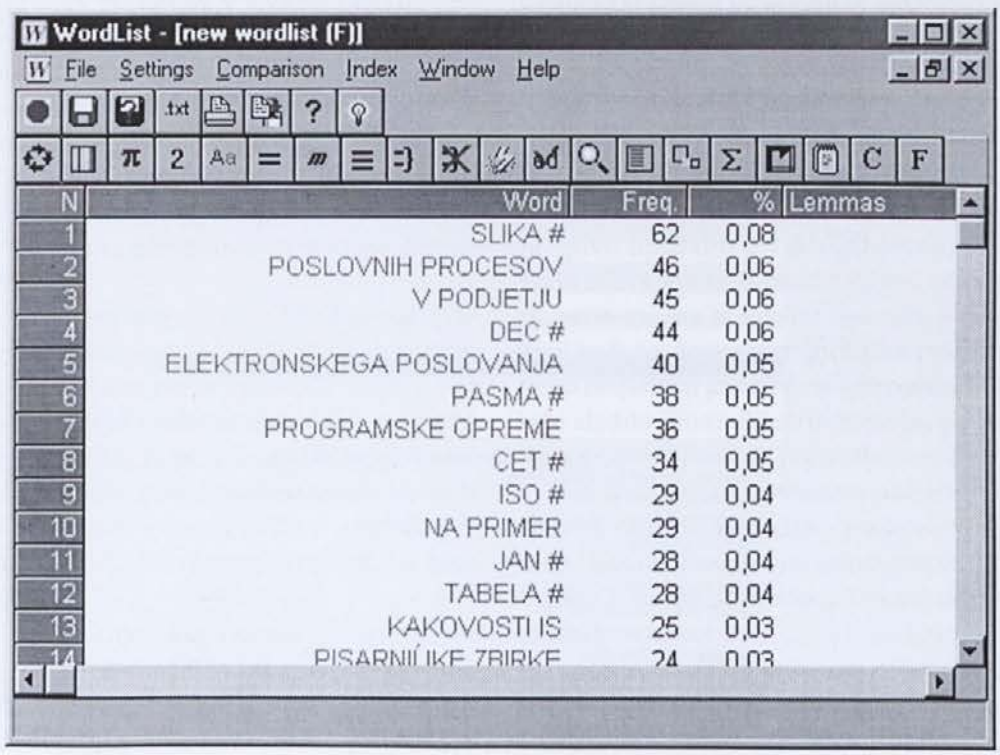
Pri zbiranju gesel si torej lahko precej pomagamo z različnimi besednimi seznami eno- in večbesednih enot ter ključnih besed. V našem primeru sta bila poglavitna kriterija za izbiro gesel pogostost in pa preverba, da geslo še ni vnešeno v spletni iSlovar. Slika 6 kaže izsek seznama pogostih dvobesednih enot v programu Wordsmith.

Program je sposoben obdelovati tudi besedila v zapisu XML ali SGML. Morda je še največja pomanjkljivost programa, da ne zna zadovoljivo obdelovati vzporednih korpusov; prikaz dvojezičnih konkordanc namreč ni mogoč. Kadar imamo opravka z dvema jezikoma, nam lahko Wordsmith pomaga le pri statistični primerjavi leksikalne gostote, povprečne dolžine odstavkov, stavkov in besed v posameznem jeziku.

Za iskanje po vzporednih korpusih so sicer tudi na voljo različna orodja, vendar je bil v našem primeru vzporedni korpus le pomožni vir izrazja, obenem pa je bilo zaradi njegove majhnosti možno po njem iskati tudi na »enojezični način«, se pravi s prikazom iskanega niza in prevedenega segmenta v isti vrstici. Za enostavno dvojezično iskanje je primeren tudi prej omenjeni program DejaVu.

4.2 Termini, razlage in kolokacije

Čeprav je informatično izrazje večinoma angleškega izvora in je privzeta smer iSlovarja angleško-slovenska, smo v našem primeru zaradi sestave korpusa izhajali iz slovenskih iztočnic, ki smo jim v drugi fazi iskali ustreznice. Ne glede na dobro računalniško podporo je izbor gesel še vedno najbolj težavna in potencialno sporna naloga v slovaropisju. Kot večina področij je



N	Word	Freq.	%	Lemmas
1	SLIKA #	62	0,08	
2	POSLOVNIH PROCESOV	46	0,06	
3	V PODJETJU	45	0,06	
4	DEC #	44	0,06	
5	ELEKTRONSKEGA POSLOVANJA	40	0,05	
6	PASMA #	38	0,05	
7	PROGRAMSKE OPREME	36	0,05	
8	CET #	34	0,05	
9	ISO #	29	0,04	
10	NA PRIMER	29	0,04	
11	JAN #	28	0,04	
12	TABELA #	28	0,04	
13	KAKOVOSTI IS	25	0,03	
14	PISARNIŠKE ZBIRKE	24	0,03	

Slika 6: **Orodje Wordsmith**

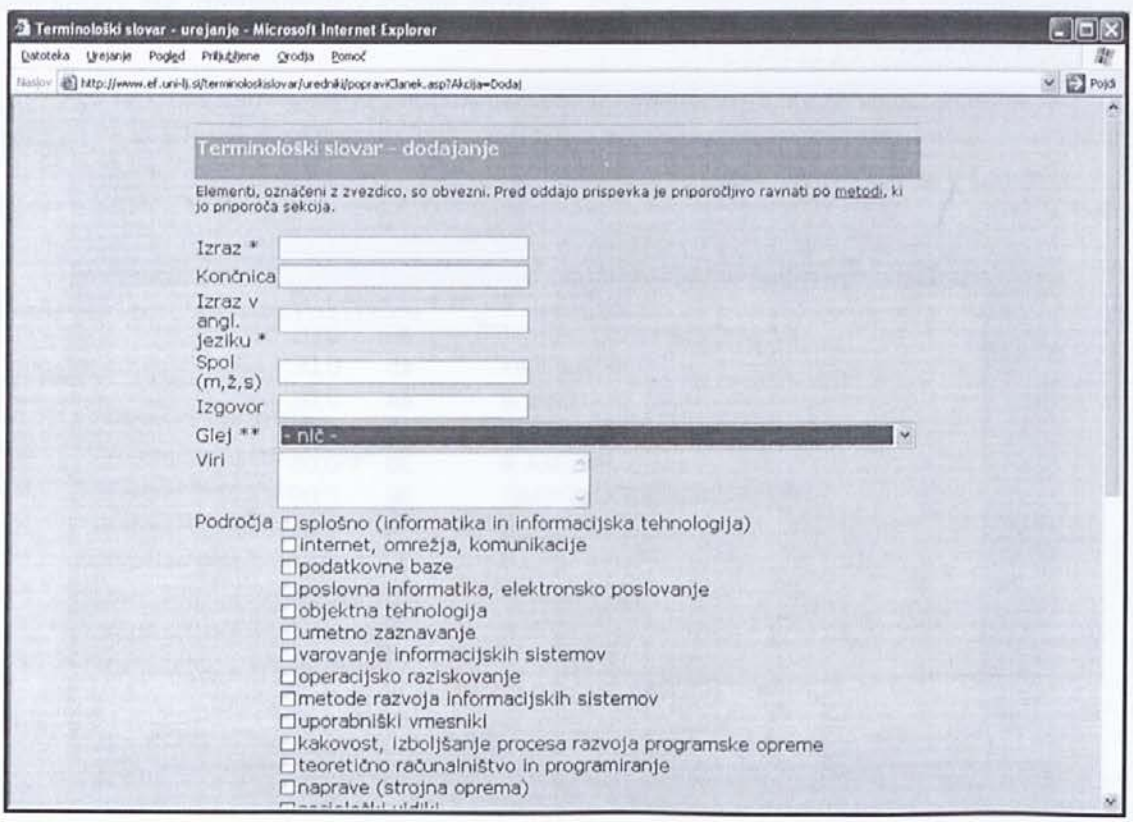
namreč tudi informatika interdisciplinarna, tako da v njej srečujemo gostujoče termine s številnih področij. Na posvetu DSI 2003 je bilo precej prispevkov namenjenih e-poslovanju, zato so bili izrazi s področja (e-)ekonomije zelo pogosti, na primer *poslovni proces, poslovni sistem, poslovna aplikacija* itd.

Študenti so imeli precej težav pri razlikovanju med terminološkimi kolokacijami, ki so se pojavljale pri vrhu Wordsmithovih besednih seznamov, in pravimi termini. Tako se na primer pojavi izraz *language technology application*, ki je verjetno kompozitivna kolokacija, kjer se pomen sestavi iz *language technology* in *application*. Čeprav smo v začetku izhajali iz načela, da bomo nediskriminatorno med gesla uvrščali tudi glagolsko izrazje in druge nesamostalniške zveze, se kmalu pokaže, da se nam glagoli kljub pogostosti in specifičnemu pomenu mnogokrat ne zdijo primerni za uvrstitev med iztočnice. Vsaj tisti, ki najbolj odstopajo od svojega splošnojezikovnega pomena, na primer *shraniti, brskati*, bi si zagotovo zaslužili terminološko obdelavo.

Med izrazi, ki so jih študenti izbrali za uvrstitev v bazo, so se znašli tudi precej splošni izrazi, ob katerih se postavlja vprašanje meje med splošnih in strokovnim besediščem, na primer *declaration – deklaracija, communication channel – komunikacijski kanal* itd. Prvi je denimo razložen kot *del uvoda kakega dokumenta*, kar je za pomensko umestitev v svet informatike zagotovo premalo, vendar bi z bolj računalniško razlago izraz lahko utemeljeno uvrstili med iztočnice in mu dodali bolj specifične podpomenke, na primer *deklaracija spremenljivk*.

Ko je bil izbor gesel končan, se je začela obdelava, se pravi opremljanje iztočnic s čim bogatejšimi slovarskimi podatki. Tu smo se navezali na obstoječo strukturo iSlovarja, ki pod posamezno iztočnico predvidi podatkovna polja izraz, končnica, izraz v angleškem jeziku, spol, izgovor, glej, viri, področje, razlaga, sinonimi in viri. Masko za spletno vnašanje, ki je dostopna le urednikom slovarja, prikazuje slika 7.

Zgornja shema ima sicer na voljo precej polj, vendar nekaterih podatkovnih kategorij ne ločuje dovolj jasno. Tako niti iz sheme niti iz pojasnila metode na spletni strani ni razvidna razlika med istoimenskima poljema Viri in razlika med polji Glej in Sinonimi. Na spletni strani je sicer podana opomba, da se pod Glej



Slika 7: **Maska za vnašanje gesel v iSlovar**

vnaša sinonime, ki imajo zaradi pomembnosti samostojen vnos. Dobra stran polja Glej je tudi, da se pojavi spustni meni, ki navaja vse že vnešene izraze. Pojavlja pa se vprašanje, kam – če sploh – je možno vnašati sorodne izraze, ki nikakor niso sinonimni, dajejo pa vseeno pomembne podatke o izrazu in sorodnih pojmih (npr. *information security – varnost podatkov; information system security – varnost informacijskega sistema).*

Najpomembnejši del obdelave gesla je iskanje razlage, ki naj bi jedrnato in hkrati dovolj natančno opredeljevala pomen iztočnice. iSlovar je namenjen predvsem slovenskim uporabnikom, zato so tudi razlage izključno slovenske. Študenti so si pri iskanju angleških razlag pomagali z različnimi viri, med drugim s funkcijo *define* v iskalniku Google in obstoječimi spletnimi slovarji informatike na internetu, najdene razlage pa so prevedli v slovenščino. Izjemoma je bilo ustrezno slovensko razlago možno najti na slovenskem spletu.

Čeprav je metoda prevajanja angleških razlag vse prej kot idealna, je vseeno vsaj osnova za oblikovanje končne različice razlage, saj smo se vseskozi zavedali, da bodo zbrana gesla vsekakor morali pregledati še strokovnjaki. Tako so razlage večinoma precej splošne in nenatančne, saj so študentje izbirali takšne, ki so jih sami razumeli. Nekaj primerov navajamo spodaj:

- *dekripcija*: vrnitev podatkov v izvirno obliko
- *digitalno potrdilo*: elektronski dokument, ki potrjuje verodostojnost osebe pri poslovanju oziroma drugih elektronskih transakcijah, na katerem je ime uporabnika, veljavnost, digitalni podpis pooblaščene osebe, ki je dokument izdala itd.
- *aplikacija jezikovnih tehnologij*: računalniški model za prepoznavanje in razumevanje naravnega človeškega govora

Skupno so študenti izbrali in obdelali okrog dvesto novih izrazov iz korpusa, poleg tega pa so z razlagami oskrbeli še približno sto že vnesenih izrazov. Ker v času pisanja še nismo imeli na razpolago povratne informacije urednikov, tega izdelka še ne moremo kakovostno ovrednotiti. Zavedati pa se moramo, da bi bilo s korpusno metodo v enakem času mogoče pridobiti tudi precej večje število izrazov, vendar brez slovarske obdelave.

S stališča prevajalcev kot rednih uporabnikov večjezičnih terminoloških del je pri večini obstoječih slovarjev, in iSlovar tu ni izjema, zanemarjena frazeološka plat strokovnega jezika. Delno se ta problem sicer rešuje s povezavo med slovarjem in korpusom, podobno zapolnitev te vrzeli sta udejanila projekta Evrokorpus in Evroterm za področje prava in evropske zakonodaje (Željko 2002). Pa vendar nam korpusne metode pri izdelavi slovarjev omogočajo lep vpogled v kolokacije in frazeologijo ob iztočnicah, česar pa se trenutno ne da umestiti v strukturo iSlovarja.

4.3 Nadaljnje jezikovne obdelave

Za izgradnjo terminološkega slovarja bi bilo seveda koristno, če bi lahko termine identificirali kar avtomatsko (Vintar, 2002). Metode samodejnega luščenja izrazov se v svetu razvijajo že nekaj časa, sprva predvsem za namene iskanja podatkov in samodejne klasifikacije dokumentov, danes pa se tovrstna orodja vgrajujejo tudi v prevajalske sisteme, kot je Tradosov. Identifikacija terminoloških kandidatov lahko temelji na statističnih metodah, ki kot kriterij upoštevajo relativno pogostost izraza in njegovo distribucijo po besedilu ali korpusu, jezikovno odvisne metode pa uporabljajo jezikoslovno analizirana besedila in izraze prepoznavajo na podlagi določenih oblikoskladenjskih vzorcev. Če so korpusi dovolj veliki, lahko najdenim izrazom statistično poiščemo tudi prevod, in tako nastaja dvojezični slovar terminoloških kandidatov brez človekove pomoči.

Seveda so zaenkrat, vsaj za slovenščino, tako pridobljeni seznami uporabni le kot osnova za nadaljnje slovarske obdelave, vendar je za številna področja, ki jim »ročno« slovaropisje ne uspe slediti, to vseeno boljše kot nič.

5 SKLEP

V članku smo predstavili izdelavo korpusa DSI ter njegovo uporabo pri dopolnjevanju spletnega slovarja SDI. Čeprav je predstavljeni postopek ilustriran na primeru, pa bi takšna uporaba jezikovnih tehnologij tudi za druga terminološka področja lahko bistveno olajšala in pohitrila slovaropisje v Sloveniji, da bi laže sledilo dinamiki strokovnega, pa tudi splošnega jezika.

Na tem mestu je vseeno potrebno opozorilo, ki zadeva celotno korpusno jezikoslovje: na korpusih temelječe analize in viri samo povzemajo jezik, ki se nahaja v korpusu. Metoda je torej deskriptivna in ne preskriptivna oz. z drugimi besedami, če so v korpusu uporabljani zastareli in neustrezni termini, bodo takšni tudi v konkordancah oz. v avtomatsko generiranem terminološkem slovarju. Za vire, ki naj bi imeli normativno funkcijo, je zato naknadna redakcija nujna.

Prihodnje delo je bilo, kar se tiče bolj bogatega označevanja, že nakazano v predhodnih poglavjih. Želeli pa bi si seveda korpus tudi razširiti. V kratkem bomo vanj dodali zbornik DSI za leto 2004, dolgoročnejši načrti pa predvidevajo zajem revije Uporabna informatika in tudi virov, ki ne izhajajo iz SDI – tu imamo v mislih predvsem vladne publikacije s področja informatike.

Izdelava korpusov in drugih jezikovnih virov je predraga, da bi bilo smiselno že v prvi fazi prepustiti njihov nastanek ekonomskim faktorjem, še posebej za jezike s tako majhnim številom govorcev, kot jih ima slovenski jezik. Z vladnim financiranjem in sodelovanjem akademskih institucij, društev, lahko pa tudi komercialnih partnerjev, kot so založbe, je nujno najprej omogočiti izdelavo predkompetitivnih virov, saj šele ti lahko dajo eno od prepotrebnih osnov za nadalnji razvoj raziskovanja in uporabe slovenskega jezika. Ti viri bi morali biti čim širše dostopni, kar lahko dosežemo po eni strani z uporabo mednarodnih standardov pri njihovem zapisu in označevanju, po drugi strani pa s čim bolj liberalnimi pogoji nadaljnjega razširjanja in uporabe. Korpus DSI je prosto dostopen za iskanje, za prepis pa ga nameravamo narediti dostopnega za raziskovalne in pedagoške namene.

Zahvala

Pri študentskem projektu so sodelovali Božo Borčnik, Katja Veber, Patricija Mencingar, Irena Perne, Jasna Čretnik, Teja Mlakar, Simona Vučak, Karmen Žerdin, Anja Čibej, Jana Štupnikar, Nina Mali, Barbara Damiš.

6 LITERATURA

[1] DŽEROSKI, Sašo, ERJAVEC, Tomaž, ZAVREL, Jakub: Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. V: *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC'00*, Athens, str. 1099–1104, 2000. http://nl.ijs.si/et/Bib/LREC00/lrec-tag-www/ SINCLAIR, John. *Looking Up: An Account of the COBUILD Project in Lexical Computing.* Collins, Glasgow. 1987.

[2] ERJAVEC, Tomaž. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V *Fourth International Conference on Language Resources and Evaluation, LREC'04*. Paris: ELRA. 2004. http://nl.ijs.si/ME/

[3] MANNING, Chirstoper, SCHÜTZE, Heinrich: *Foundations of Statistical Natural Language Processing.* The MIT Press. Cambridge MA. 1999.

[4] SPERBERG-MCQUEEN, C.M.; BURNARD, Lou (ur.). *Guidelines for Electronic Text Encoding and Interchange, the XML Version*. TEI Consortium, 2002. http://www.tei-c.org/

[5] VAN HALTEREN, Hans (ur.) *Syntactic Wordclass Tagging*. Kluwer, 1999.

[6] VINTAR, Špela: Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil. V: *Zbornik 3. konference o jezikovnih tehnologijah*, Ljubljana, str. 78–85, 2002. http://nl.ijs.si/isjt02/zbornik/sdjt02-14vintar.pdf

[7] W3C. *Extensible Markup Language (XML) 1.0 (Second Edition)*. (2000). http://www.w3.org/XML/

[8] W3C. *XSL Transformations (XSLT) Version 1.0 (1999)* http://www.w3.org/TR/xslt

[9] ŽELJKO, Miran: Pripomočki na spletu za prevajalce zakonodaje EU. V: *Zbornik 3. konference o jezikovnih tehnologijah*, Ljubljana, str. 33–39, 2002. http://nl.ijs.si/isjt02/zbornik/sdjt02-05zeljko.pdf

Dr. Tomaž Erjavec je znanstveni sodelavec na Odseku za tehnologije znanja na Institutu Jožef Stefan. Njegov raziskovalni interes je računalnišo jezikoslovje, tj. jezikovne tehnologije, korpusno jezikoslovje in strojno prevajanje, predvsem v povezavi s slovenskim jezikom. Diplomiral je na Fakulteti za elektrotehniko in računalništvo Univerze v Ljubljani (1984), magistiral pa na Fakulteti za računalništvo in informatiko (1990) in na Centre for Cognitive Science Univerze v Edinburghu (1992), doktoriral je na Fakulteti za računalništvo in informatiko Univerze v Ljubljani (1997). Je avtor več kot 50 znanstvenih člankov, član uredniških odborov mednarodnih revij CHum in IJCL, predsednik slovenskega društva za jezikovne tehnologije, bil pa je tudi član sveta Text Encoding Initiative Consortium ter European Chapter of the Association of Computational Linguistics.

Špela Vintar je na Filozofski fakulteti diplomirala iz angleščine in nemščine, zatem pa v okviru podiplomskega študija preživela nekaj večmesečnih obdobij na raziskovalnih projektih v tujini. Leta 2003 je doktorirala s področja samodejnega luščenja terminologije iz slovenskih in angleških besedil. Od leta 1998 je zaposlena na Oddelku za prevajalstvo Filozofske fakultete. Od tedaj se ves čas intenzivno ukvarja s korpusi, v zadnjem času pa tudi s poučevanjem korpusnih metod v prevajalstvu in slovaropisju.