KORPUSI IN KONKORDANČNIKI NA STREŽNIKU NL.IJS.SI

Tomaž ERJAVEC
Institut "Jožef Stefan", Odsek za tehnologije znanja

Erjavec, T. (2013): Korpusi in konkordančniki na strežniku nl.ijs.si. Slovenščina 2.0, 1 (1): 24–49.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

V prispevku predstavimo referenčne, specializirane in vzporedne korpuse, do katerih je mogoče dostopati prek konkordančnikov na strežniku nl.ijs.si. Večina korpusov vsebuje besedila v slovenščini, nekaj pa je tudi tujejezičnih. Mnogi od korpusov obstajajo že dalj časa, vendar so sedaj na novo označeni, pri nekaterih so dodana nova besedila, nekateri pa so v celoti novi. Besedila v korpusih so opremljena z metapodatki, besednim pojavnicam pa so ročno ali avtomatsko pripisane vsaj leme in oblikoskladenjske oznake. V večini primerov so korpusi prosto dostopni, in sicer prek dveh spletnih konkordančnikov, ki omogočata iskanje po obsežnih označenih korpusih, ponujata bogat nabor analitičnih orodij, možnosti filtriranja glede na metapodatke in shranjevanje rezultatov na lastni računalnik. Poleg korpusov in obeh konkordančnikov v prispevku obravnavamo tudi nekatera vprašanja, ki so se zastavila pri zagotavljanju tovrstne infrastrukture za namene korpusnega jezikoslovja, ter zaključimo s smernicami za nadaljnje delo.

Ključne besede: računalniški korpusi, konkordančniki, CWB, CUWI, noSketchEngine

1 UVOD

Za slovenski jezik obstaja na spletu več referenčnih korpusov, kot so na primer Nova Beseda (Jakopin in Michelizza 2007), FidaPLUS (Arhar Holdt in Gorjanc 2007) in najnovejši Gigafida (Logar Berginc in dr. 2012). Zadnji je,

kot še nekateri drugi korpusi, rezultat projekta Sporazumevanje v slovenskem jeziku (SSJ).[1] Že dlje časa so na strežniku nl.ijs.si dostopni tudi angleško-slovenski vzporedni korpus evropskega pravnega reda IJS-ELAN (Erjavec 2002), korpus informacijskega izrazja DSI (Puc in Erjavec 2006), spletni enojezični in vzporedni korpus japonskega jezika, ki sta bila izdelana kot podpora slovarju za slovenske učence japonščine jaSlo (Hmeljak Sangawa in Erjavec 2008), ter referenčna korpusa jos100k in jos1M, izdelana pri projektu JOS[2] (Erjavec in dr. 2010).

Korpusi, ki jih najdemo na strežniku, so bili izdelani v obdobju več kot petnajstih let. Kodirani in jezikoslovno označeni so bili sicer v skladu s Priporočili za zapis besedil TEI[3] (Text Encoding Initiative), vendar so uporabljali vse različice priporočil, od TEI P3 iz leta 1994, pa do zadnje, P5, iz leta 2007. Neusklajen je bil tudi dostop do korpusov prek konkordančnika na strežniku, saj je imel skoraj vsak korpus drugačen spletni vmesnik, zaledni del pa je bil zelo stara različica prosto dostopnega iskalnika po korpusnih podatkih CWB (Christ 1994). Strežnik tako ni podpiral enotnega iskanja po korpusih, vmesniki pa tudi niso ponujali naprednih funkcionalnosti, kot npr. prikaz in filtriranje po metapodatkih, sortiranje, izračun kolokacij ali shranjevanje rezultatov na lastni računalnik.

Osnovni namen članka je predstaviti korpuse, ki so namenjeni jezikoslovnim raziskavam. Najprej v drugem razdelku predstavimo posodobljeno zbirko korpusov, ki so sedaj enotno označeni po TEI P5, nekaterim so dodana nova besedila, nekateri korpusi pa so povsem novi. Tretji razdelek je namenjen predstavitvi obeh sodobnih in visokozmogljivih konkordančnikov, ki omogočata dostop do naštetih korpusov, v četrtem razdelku pa podamo zaključke in smernice za nadaljnje delo.

[1] http://www.slovenscina.eu/

[2] http://nl.ijs.si/jos/

[3] http://www.tei-c.org/

2 KORPUSI

V tem razdelku najprej na kratko predstavimo korpusne oznake, nato pa naštejemo korpuse, ki so trenutno dostopni prek konkordančnika. Pregled posameznih korpusov s kvantitativnimi kazalci je podan v Tabeli 1.

ime (pod)korpusa*	jezik	zvrst	milijonov pojavnic	milijonov besed	besedilnih enot	od leta	do leta
Gigafida	sl	C/A	1.409,76	1.187,00	39.427	1990	2011
⊃ KRES	sl	V/A	120,45	99,83	21.456	1990	2011
⊃ ssj500k	sl	V/R	0,59	0,50	1.655	1994	2006
GOS	sl	C/A	1,11	1,11	287	2004	2010
IMP	sl	C/A	8,80	7,16	29.227	1584	1918
⊃ goo300k	sl	V/R	0,30	0,24	81	1584	1899
VAYNA	sl	C/A	0,26	0,23	355	1987	1988
Tweet-sl	sl	V/A	6,29	5,02	367.51	2007	2011
DSI	sl	C/A	3,49	2,98	1.365	2003	2012
SDJT	sl	C/A	0,33	0,28	183	1998	2010
FILMI	sl	C/A	0,94	0,78	2.449	2000	2009
KONJI	sl	C/A	0,47	0,40	436	1995	2008
TRANS5	en	C/A	1,83	1,60	128	1984	2012
TRANS5	sl	C/A	1,59	1,35	128	1983	2012
⊃ slRev	sl	C/A	0,15	0,12	25	2006	2006
⊃ slRev	en	C/A	0,18	0,15	25	2006	2006
SPOOK-de	de	C/A	0,55	0,47	7	1983	2008
SPOOK-de	sl	C/A	0,53	0,44	7	2000	2010
SPOOK-en	en	C/A	1,16	0,97	9	1992	2007
SPOOK-en	sl	C/A	1,15	0,92	9	2002	2008
SPOOK-fr	fr	C/A	0,81	0,70	12	1989	2006
SPOOK-fr	sl	C/A	0,72	0,60	12	1997	2008
SPOOK-it	it	C/A	0,49	0,41	7	1992	2001
SPOOK-it	sl	C/A	0,48	0,39	7	1999	2004
SPOOK-sl	sl	C/A	1,64	1,34	25	1996	2006

LeMonde	fr	C/A	0,72	0,63	300	2005	2009
LeMonde	sl	C/A	0,62	0,52	300	2006	2009
EU DGT	sl	C/A	34,32	28,92	19.661	2004	2011
EU DGT	en	C/A	37,80	33,36	19.661	2004	2011
EU DGT	de	C/A	34,67	30,24	19.661	2004	2011
EU DGT	fr	C/A	42,27	37,86	19.661	2004	2011
EU DGT	it	C/A	39,86	35,18	19.661	2004	2011
jaSlo	jp	C/A	0,76	0,65	132	1916	2009
jaSlo	sl	C/A	0,53	0,44	132	1920	2009
jpWaC-L	jp	C/A	409,03	325,66	49.536	?	2007
⊃ jpWaC-L0	jp	V/A	43,76	39,27	46.461	?	2007
⊃ jpWaC-L1	jp	V/A	1,63	1,45	31.233	?	2007
⊃ jpWaC-L2	jp	V/A	4,61	4,09	37.365	?	2007
⊃ jpWaC-L3	jp	V/A	1,04	0,91	23.314	?	2007
⊃ jpWaC-L4	jp	V/A	0,30	0,26	14.316	?	2007
ELIZA	en	V/A	24,99	22,35	5.728.568	2002	2007
Σ 41 (⊅ 31)	6	–	2.067,97	1.729,95	6.318.863	1584	2012

Tabela 1: Osnovni podatki o korpusih, vključenih v konkordančnika na strežniku nl.ijs.si.

* Stolpci vsebujejo: ime korpusa oz. vsebovanega podkorpusa; kodo jezika korpusa (istoimenski so poravnani), zvrst (celotna besedila (C) oz. vzorčena besedila (V) in avtomatsko (A) oz. ročno označena besedila (R)), število milijonov pojavnic in besed, število vsebovanih besedilnih enot ter časovno obdobje besedil iz korpusa. Kjer podatka ni, je podan vprašaj. Sumarna vrstica vsebuje število korpusov, če (ne) štejemo podkorpusov, število različnih jezikov v korpusih, število pojavnic in besed neprekrivnih korpusov ter letnico najstarejšega in najmlajšega v korpuse vključenega besedila.

2.1 Korpusne oznake

Korpusi vsebujejo podatke o strukturi zajetih besedil, kot so besedilo, odstavek ali stran. Strukturnim oznakam so pridruženi metapodatki, npr.

kazalka na faksimile posamezne strani, za besedilo pa naslov, avtor, leto izdaje, založba itd.

Korpusi na strežniku so tudi jezikoslovno označeni. Prva stopnja označevanja, ki je potrebna že za delovanje konkordančnika, je tokenizacija, tj. razdelitev besedila na besede in ločila, v veliki večini korpusov pa so označeni tudi presledki ter povedi. Besednim pojavnicam sta nadalje pripisani vsaj še oblikoskladenjska oznaka in lema. Oblikoskladenjske oznake večinoma sledijo priporočilom za oblikoskladenjsko označevanje JOS[4] (Erjavec in Krek 2008).

Korpusi so pretežno avtomatsko označeni, pri nekaterih referenčnih korpusih pa je jezikoslovno označevanje potekalo ročno. Za avtomatsko označevanje je bil v večini primerov uporabljen program ToTaLe (Erjavec in dr. 2005), ki opravi *to*kenizacijo, oblikoskladenjsko označevanje (*ta*giranje) in *le*matizacijo. Pri korpusih, ki so bili označeni na drug način, to v nadaljevanju posebej izpostavimo.

2.2 Referenčni korpusi slovenskega jezika

Pri dveh sklopih korpusov slovenskega jezika je bila že v izhodišču izpostavljena težnja po reprezentativnosti, zato vsebujejo veliko raznovrstnih besedil. Gigafida in njene izpeljanke vsebujejo sodobna slovenska besedila, medtem ko segajo besedila v korpusih IMP[5] v čas pred koncem prve svetovne vojne. Poleg celotnih korpusov sta na voljo tudi vzorčena in dosti manjša, vendar ročno označena podkorpusa ssj500k iz Gigafide in goo300k iz projekta IMP, ki sta primerna predvsem za šolanje avtomatskih označevalnikov in za jezikoslovne študije, pri katerih je zaželeno, da so oznake čim bolj pravilne.

2.2.1 REFERENČNI KORPUSI SODOBNE SLOVENŠČINE PROJEKTA SSJ

Največji korpus na strežniku je referenčni korpus sodobnega (1990–2011)

4 http://nl.ijs.si/jos/josMSD-sl.html

5 IMP je okrajšano ime projekta EU IMPACT »Improving Access to Text«, ki je v veliki meri podprl naše delo na starejši slovenščini.

slovenskega jezika **Gigafida** (Logar Berginc in dr. 2012) z več kot milijardo besed. Besede v korpusu so lematizirane in imajo avtomatsko pripisane oblikoskladenjske oznake, za kar je bil uporabljen označevalnik Obeliks[6] (Grčar in dr. 2012).

Desetkrat manjši uravnoteženi korpus **KRES** (*ibid*) s sto milijoni besed je po odstavkih vzročen iz Gigafide in je nastal z namenom, da se v njem uravnotežijo besedilne zvrsti, zato naj bi bili npr. frekvenčni seznami besed bolj v skladu s »tipično« slovenščino.

Korpus **ssj500k** (Holozan in dr. 2008) ima pol milijona besed in vsebuje ročno označen korpus jos100k in štiristo tisoč besed iz delno ročno označenega jos1M (Erjavec in dr. 2010). Za ssj500k so bile oznake v celoti ročno preverjene (Arhar 2009), delno je tudi ročno označen s površinskoskladenjskimi povezavami, ki pa jih konkordančnik še ne podpira.

Korpus govorjene slovenščine **GOS** (Verdonik in Zwitter Vitez 2011) vsebuje nekaj več kot milijon besed. Različica korpusa na strežniku nl.ijs.si sicer ne vsebujejo govora (te datoteke niso javno dostopne), pač pa transkripcije besedil, pri katerih je vsaki besedi pripisana tudi njena normalizirana oblika, kar olajša iskanje in prikaz besed v korpusu.

Gigafida, KRES in GOS so izvorno dostopni prek konkordančnikov projekta SSJ[7] in so, poleg korpusov DGT, zaenkrat tudi edini v tem prispevku opisani korpusi, do katerih lahko dostopamo in jih uporabljamo tudi prek drugega spletnega konkordančnika.

2.2.2 REFERENČNI KORPUSI STAREJŠE SLOVENŠČINE IMP

Korpus starejše slovenščine **IMP**[8] (Erjavec 2012) vsebuje večinoma ročno pregledana (iz skenogramov zajeta) besedila celotnih knjig iz obdobja 1750–

[6] http://sourceforge.net/projects/obeliks/ oz.
http://www.slovenscina.eu/tehnologije/oznacevalnik
[7] http://www.gigafida.net/, http://www.korpus-kres.net/ in http://www.korpus-gos.net/
[8] http://nl.ijs.si/imp/

1918, kot tudi nekaj rokopisov ter izdaj časopisa Kmetijske in rokodelske novice, skupaj več kot sedem milijonov besed. Vsaka stran besedila v korpusu je povezana s faksimilom in ustrezno stranjo v digitalni knjižnici IMP. Korpus je z vidika jezikoslovnega označevanja poseben, saj je besednim pojavnicam poleg (sodobne) leme pripisana tudi sodobna ustreznica besedne oblike. Označen je avtomatsko s programom ToTrTaLe (Erjavec 2011), ki poleg tokenizacije, oblikoskladenjskega označevana in lematizacije opravi tudi *tr*anskripcijo, tj. posodobi zastarele besedne oblike.

Referenčni korpus **goo300k** vsebuje tisoč strani (okoli tristo tisoč pojavnic), ki so bile vzorčene iz korpusa IMP. Transkripcije so bile dodatno ročno popravljene, predvsem pa je bil korpus ročno jezikoslovno označen. V primerih, ko gre za zastarelo besedo, so bili poleg ročno preverjene sodobne besedne oblike in leme besedi pripisani tudi najbližji sodobni sinonimi oz. razlaga. Ker je bil poudarek pri označevanju na posodabljanju besed in ne na oblikoskladenjskem označevanju, so oblikoskladenjske oznake tu manj podrobne kot pri korpusih sodobne slovenščine. Oblikoskladenjske specifikacije IMP[9] ne zajemajo pregibnih lastnosti, kot sta sklon ali oseba, poenostavijo pa tudi leksikalne lastnosti, tako da npr. zaimke zastopa ena sama oznaka. Nabor oznak se je pri tem korpusu tako zmanjšal s skoraj dva tisoč, kolikor jih definirajo specifikacije JOS, na dvaintrideset.

2.3 Drugi enojezični korpusi slovenščine

Strežnik trenutno ponuja šest enojezičnih specializiranih korpusov slovenskega jezika, torej korpusov, ki so bolj ozko osredotočeni na določen podjezik, posamezno temo ali besedilni tip. Ti korpusi so bili zgrajeni za namene določene raziskave, za izdelavo slovarja oz. kot učni pripomočki ali pa so priložnostni: če so bili na voljo zanimivi jezikovni podatki, smo jih pretvorili v korpus z željo, da bodo uporabni za jezikoslovne ali jezikovnotehnološke raziskave.

[9] http://nl.ijs.si/imp/msd/html-sl/

2.3.1 NOVEJŠA SLOVENSKA ZGODOVINA

Korpus **VAYNA** oz. Korpus verbalnih napadov na JNA je najstarejši korpus, ki je bil na Institutu "Jožef Stefan" izdelan z namenom objektivno ovreči tezo, da se v slovenskem tisku napada Jugoslovansko narodno armijo (Tancig in Žagar 1989). Korpus vsebuje 360 člankov (220 tisoč besed), ki so bili v letih med 1987 in 1988 objavljeni v periodiki, kot so Delo, Dnevnik, Komunist, Mladina, Teleks itd. Korpus vsebuje predvsem komentarje slovenskega novinarstva na dogodke v obdobju t. i. slovenske pomladi, ki jo je najbolj zaznamoval proces proti četverici JBTZ.[10]

2.3.2 DRUŽABNO OMREŽJE TWITTER

Družabna omrežja so v jezik prinesla veliko novosti, saj vsebujejo tipično neformalno in neknjižno izrazje, uvajajo nove kratice ipd. Med bolj znanimi mediji je Twitter, na katerem se objavljajo »*tweeti*« oz. tviti, tj. kratka sporočila, ki jih vidijo vsi sledilci avtorja oz. teme. Zaradi priljubljenosti tvitov so se pojavili tudi agregatorji, ki jih zbirajo in analizirajo ter ponujajo tiste, ki naj bi bili najbolj zanimivi za določen profil uporabnikov. Z agregatorja *sitweet.com* so nam prijazno odstopili bazo slovenskih tvitov, ki so nastali med letoma 2007–2011. Bazo smo prečistili, odstranili smo npr. reklame (*spam*) in tujejezične tvite, ter tako dobili korpus **Tweet-sl**, ki vsebuje 360.000 tvitov oz. pet milijonov besed. Korpus je sicer avtomatsko jezikoslovno označen, vendar pa je natančnost nizka, saj so bili uporabljeni modeli naučeni na virih standardne slovenščine, metode za normalizacijo »tvit slovenščine« pa bo potrebno šele razviti.

2.3.3 INFORMATIKA IN JEZIKOVNE TEHNOLOGIJE

Korpus **DSI** (Puc in Erjavec 2006) se oblikuje predvsem kot podpora spletnemu terminološkemu slovarju informatike iSlovar.[11] Trenutno vsebuje devet zbornikov konference Dnevi slovenske informatike (2003–2012) in

[10] http://sl.wikipedia.org/wiki/proces_proti_%c4%8detverici

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

izbrane številke (2010–2012) revije Uporabna informatika, kar je skupaj več kot 1.300 člankov in skoraj tri milijone besed.

Korpus **SDJT** vsebuje slovenske prispevke iz zbornikov konferenc Jezikovne tehnologije (1998–2010), ki potekajo vsako drugo leto v sklopu metakonference Informacijska družba. Zborniki so dostopni na spletu v formatu PDF, iz katerih je bil izdelan besedilni korpus (Smailović in Pollak 2011), njegov slovenski del pa tvori korpus SDJT, ki vsebuje 183 člankov oz. 280.000 besed.

2.3.4 KONJENIŠTVO IN FILMSKE KRITIKE

Zadnja dva specializirana korpusa sta nastala kot del magistrskih nalog, v sklopu katerih sta bila poleg korpusov izdelana tudi terminološka slovarčka, in sicer konjeništva (Plahuta 2010) in filmskih kritik (Košir 2010). Korpus **KONJI** vsebuje tri knjige in izvode dveh revij, ki se ukvarjajo s konjeništvom, in ima skupno 436 enot ter 400.000 besed. Korpus **FILMI** vsebuje recenzije filmov, ki so bile objavljene v revijah Mladina in Premiera ter v prilogi časopisa Delo Vikend, skupno gre za skoraj 2.500 člankov oz. 777.000 besed.

2.4 Vzporedni korpusi

Vzporedni korpusi vsebujejo besedila s stavčno poravnanimi prevodi, pri čemer je trenutno na konkordančnikih eden od jezikov vedno slovenščina. Taki korpusi so zelo koristni za prevajalce, nujno potrebni pa so tudi za šolanje ali testiranje strojnih prevajalnikov. Izdelava tovrstnih korpusov je glede na enojezične veliko bolj zapletena, saj je težko pridobiti digitalno predlogo besedila in prevoda, poravnavo je za zadovoljivo kvaliteto potrebno opraviti ročno, jezikoslovno pa je dobro označiti oba jezika s skladnimi oznakami.

Kjer ni drugače navedeno, smo spodaj naštete korpuse označili z metodo, ki je bila razvita in uporabljena za označevanje korpusa SPOOK (Erjavec 2013). Za slovenska besedila smo uporabili ToTaLe, za tujejezične dele pa označevalnik

in lematizator TreeTagger[12] (Schmid 1994). TreeTaggerjeve oznake, ki so si med jeziki zelo različne, so preslikane v oznake, ki so skladne z oblikoskladenjskimi specifikacijami SPOOK,[13] ki poleg slovenščine pokrivajo še angleščino, francoščino, italijanščino in nemščino.

2.4.1 ANGLEŠKO-SLOVENSKI KORPUSI

Angleški jezik je zaradi svoje razširjenosti ponavadi prvi, za katerega je izdelan vzporedni korpus za nek jezik. Že do sedaj je bilo na strežniku dostopnih več angleško-slovenskih korpusov, ki so nastajali v daljšem obdobju. Sedaj smo nekatere združili v korpus **TRANS5**, ki smo mu dodali nova besedila in ga na novo označili. Korpus ima 1.350.000 besed (v tem razdelku vedno navajamo samo obseg slovenskega dela korpusa), vanj pa so združeni naslednji korpusi:

- **Trans** (700 tisoč besed) vsebuje v daljšem obdobju priložnostno zbrana in poravnana besedila z Oddelka za prevajanje FF (Vintar in Erjavec 2000), npr. članke iz revije *Adria In-Flight Magazine*, znanstvene članke o elektrarni Krško, pa tudi nekaj leposlovja;
- **IJS-ELAN** (500 tisoč besed) je nastal v okviru evropskega projekta ELAN (Erjavec 2002) in vsebuje poravnana besedila iz dvanajstih virov, od ustave RS ter osamosvojitvenega govora M. Kučana do učbenika za delo z operacijskim sistemom Linux in Lekovega Vademecuma;
- **JRC-ECDC** (36 tisoč besed) je pomnilnik prevodov, ki je nastal na Evropskem centru za preprečevanje in nadzorovanje bolezni, in vsebuje poravnane prevodne enote s tega področja;[14]
- **slRev** (120 tisoč besed) vsebuje posebno dvojezično številko Slavistične revije iz leta 2006, ki vsebuje petindvajset jezikoslovnih prispevkov.[15]

[12] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[13] http://nl.ijs.si/spook/

[14] Pomnilnik prevodov ECDC, ki v celoti vsebuje 25 jezikov, je prosto dostopen na spletnih straneh JRC (*Joint Research Center EU*).

[15] Digitalno predlogo je zagotovil Vojko Gorjanc, stavčno poravnavo pa je opravila Darja Fišer.

> Zaradi zanimivosti je korpus dostopen v konkordančnikih tudi kot posebna enota.

TRANS5 vsebuje torej predvsem raznovrstna stvarna, večinoma terminološko bogata besedila, čeprav imajo podkorpusi tudi nekaj leposlovja, npr. Orwellov *1984* (Orwell 1949; 1984) in *Pygmalion* (Shaw 1916; 1997). Korpus tako nima posebnih jezikoslovnih vodil, v ospredju je bila predvsem želja izdelati čim bolj raznovrsten, predvsem pa čim večji vzporedni in ročno poravnani korpus.

2.4.2 VEČDVOJEZIČNI KORPUS SODOBNEGA LEPOSLOVJA SPOOK

Večkratni dvojezični korpus sodobnega leposlovja **SPOOK** je bil narejen v okviru istoimenskega projekta, katerega cilj je bil omogočiti in izvajati korpusno osnovane prevodoslovne raziskave (Vintar 2013). Korpus vsebuje izvirnike 35 romanov v angleščini, francoščini, italijanščini in nemščini ter njihove prevode v slovenščino, poleg tega pa še 25 izvirnih slovenskih romanov, skupaj štiri milijone slovenskih besed. Ker avtorske pravice za knjige ne dopuščajo prostega razširjanja, je ta korpus tudi eden redkih na strežniku, ki niso prosto dostopni.

2.4.3 FRANCOSKO-SLOVENSKI KORPUS

LeMonde (Mezeg 2010) je francosko-slovenski korpus, ki vsebuje 300 člankov (nekaj več kot pol milijona besed), objavljenih med letoma 2006–2009 v tedniku *Le Monde Diplomatique*, in njihovih prevodov v slovenščino, objavljenih v prilogi časopisa Delo. Zagotovitev digitalne predloge, pridobitev dovoljenja za uporabo v raziskovalne namene, poravnava in označevanje so bili narejeni v okviru doktorske disertacije (Mezeg 2011), korpus pa je jezikoslovno označen po metodologiji SPOOK.

2.4.4 VEČJEZIČNI KORPUS EU-DGT

Vzporedni korpus je **EU-DGT** (Steinberger in dr. 2012) vsebuje pomnilnik prevodov, ki je nastal pri prevajanju evropskega pravnega reda, in vsebuje 29

milijonov besed. Ta oz. zelo podoben korpus je že dlje časa dostopen prek spletnega iskalnika EVROKORPUS[16] (Željko in Krstič 2002), kjer je povezan tudi z večjezičnim tezavrom. Izvorno vsebuje 22 jezikov, v naš korpus pa smo vključili samo pet jezikov projekta SPOOK, saj smo imeli za te že razvit postopek avtomatskega označevanja.

2.4.5 JAPONSKO-SLOVENSKI KORPUS

Korpus **jaSlo** (Hmeljak Sangawa in Erjavec 2008) nastaja kot podprojekt izdelave istoimenskega slovarja (Erjavec in dr. 2004) in vsebuje romane, spletna besedila, predloge predavanj itd. Namen korpusa je služiti kot pripomoček za učenje japonskega jezika in kot vir dodatnih primerov za slovar jaSlo.[17] Besedila se zbirajo in poravnavajo na Oddelku za azijske in afriške študije Filozofske fakultete Univerze v Ljubljani; zadnja so bila dodana leta 2012, tako da korpus trenutno vsebuje okoli pol milijona besed iz 132 različnih virov. Japonski del korpusa smo jezikoslovno označili s programom *Chasen*[18] (Matsumoto in dr. 2003), tako kot tudi enojezični japonski korpus, ki ga predstavljamo v naslednjem poglavju.

2.5 Tujejezični korpusi

Čeprav je namen strežnika ponuditi predvsem korpuse slovenskega jezika, smo izdelali tudi dva povsem tujejezična korpusa, in sicer japonščine ter angleščine.

2.5.1 SPLETNI KORPUS ZA UČENJE JAPONŠČINE

Korpus japonskega jezika, zbran s svetovnega spleta jpWaC (Srdanović in dr. 2008), je obsežen enojezični korpus (50.000 spletnih strani, 300 milijonov besed), v katerem so bile besede in povedi označene glede na težavnostno stopnjo (Hmeljak Sangawa in dr. 2010), s čimer smo dobili spletni korpus za

[16] http://evrokorpus.gov.si/
[17] http://nl.ijs.si/jaslo/
[18] http://chasen-legacy.sourceforge.jp/

učenje japonščine kot tujega jezika **jpWaC-L2**. Konkordančnika ponujata tudi pet podkorpusov jpWaC-L2, pri čemer vsak vsebuje samo povedi določene težavnostne stopnje, namenjeni pa so kot vir primerov za učitelje oz. učence japonščine na določeni stopnji učenja.

2.5.2 POGOVORI S PROGRAMOM ELIZA

ELIZA (Weizenbaum 1966) je, vsaj v zgodovini razvoja tega področja, verjetno najbolj znan program umetne inteligence. Program je zasnovan tako, da se pogovarja z uporabnikom, pri čemer posnema psihiatra, ki pa zgolj zastavlja vprašanja. Program je presenetljivo preprost, saj je njegova naloga zgolj obrniti izjavo uporabnika v vprašanje, npr. če je v uporabnikovi izjavi beseda »*not*«, program vpraša »*Why are you so negative?*«. Program je bil implementiran v številnih programskih jezikih, mdr. v Prologu, to implementacijo pa je kmalu po začetku obstoja svetovnega spleta (pri katerem je bil IJS eno zgodnjih vozlišč) postavil na splet[19] Marko Grobelnik. Ker je bil program med prvimi na spletu in deluje že več kot petnajst let, je postal izredno priljubljen. Tako ga npr. Google postavlja na prvo mesto med 84 milijoni zadetkov za poizvedbo »*eliza*«. Spletna postavitev Elize ima izredno neinformativno masko, tako da uporabniki nimajo prave predstave, česa je sposobna in česa ne, iz česar pogosto izvira velika frustracija v komunikaciji, kar se odraža tudi v odgovorih uporabnikov na Elizina vprašanja.

Avtor pričujočega članka je program vzdrževal skozi več selitev strojne in programske opreme, hranil pa je tudi dnevnik programa, v katerem so zapisane izjave uporabnikov in vprašanja nanje, vključno s časom, ne hrani pa podatkov o spletnih naslovih, od koder je poizvedba prišla. Ta dnevnik smo sedaj nadgradili v korpus **ELIZA**, pri čemer smo prečistili šumne podatke (programske napake, napadi robotov) in ga jezikoslovno označili. Na ta način smo dobili zelo obsežen specializiran korpus, ki vsebuje skoraj šest milijonov izjav uporabnikov oz. 22 milijonov besed, zbranih v obdobju med 2002 in 2007.

[19] http://www-ai.ijs.si/eliza/

Korpus ELIZA sicer ne vsebuje osebnih podatkov uporabnikov, vendar pa je bil zgrajen na podlagi dnevnika programa, ki na spletni strani ne opozarja na to, da so podatki lahko namenjeni nadaljnjim obdelavam oz. objavam, zaradi česar je dostop do korpusa mogoč samo z geslom, ki ga lahko dobijo zainteresirani raziskovalci.

3 KONKORDANČNIKA

Predstavljeni korpusi so dostopni prek dveh konkordančnikov, CUWI (Erjavec 2013) in noSketchEngine (Rychlý 2007). Konkordančnika sta sestavljena iz čelnega dela, tj. spletnega vmesnika, ki uporabnikovo zahtevo prevede v poizvedbo, ter zalednega dela, ki na poizvedbo vrne rezultate iz izbranega korpusa. Čelni del te rezultate nato oblikuje za prikaz na zaslonu. Konkordančnika imata skupno zgodovino, saj temelji zaledni del obeh na programu CWB – *The IMS Open Corpus Workbench* (Christ 1994), ki vhodne korpuse najprej indeksira, kar omogoča hitra in kompleksna iskanja po obsežnih označenih korpusih. Računalniški jezik, ki se uporablja za poizvedbe, se imenuje CQL – *Corpus Query Language* in omogoča iskanje z regularnimi izrazi (npr. besede, ki vsebujejo »*gn*« ali so sestavljene samo iz soglasnikov), iskanje po kombinaciji oznak za posamezno pojavnico (npr. lema *pot* v ženskem spolu) in iskanje po nizih pojavnic (npr. beseda *fašističen*, ki jo loči manj kot pet besed od besede *JNA*). Program CWB obstaja že dolgo časa in se še vedno odprtokodno vzdržuje pod imenom *Open Corpus Workbench*;[20] kot visokozmogljiv sistem ga uporabljajo številni korpusi po svetu.

Zalednemu delu CWB smo v sklopu projekta SPOOK napisali lastni čelni del, po imenu **CUWI** – *Corpus Users' Web Interface* (Erjavec 2013), ki je sicer še v razvoju, vendar že ponuja različne funkcionalnosti, predvsem tiste, ki so nam koristile pri tekočih projektih, kot npr. iskanje in prikaz po skupinah (vzporednih) korpusov za SPOOK ter prikaz faksimilov ali več oznak hkrati, npr. poravnanih izvornih in posodobljenih besednih oblik za korpuse starejše

[20] http://cwb.sourceforge.net/

slovenščine IMP.

Funkcionalnost programa CWB, kot tudi formata vhodnih podatkov sta bila prevzeta, vendar na novo implementirana v programu Manatee, ob tem pa je bil razvit tudi čelni del Bonito (Rychlý 2007). Manatee in Bonito se uporabljata v sklopu komercialnega konkordančnika SketchEngine,[21] ki pa je pred kratkim dobil svojo odprtokodno različico **noSketchEngine**.[22] Ciljni uporabniki (no)SketchEnginea so jezikoslovci, predvsem leksikografi, zato je vmesnik bolj jezikoslovno usmerjen in jezikoslovcem verjetno tudi bolj intuitiven.

3.1 Iskanje in prikaz

Oba konkordančnika ponujata več načinov iskanja in prikaza rezultatov. V pričujočem razdelku jih na kratko naštejemo, bolj podrobno pa so opisani v spremni dokumentaciji konkordančnikov.

Iskanje je lahko enostavno ali pa napredno, npr. z izrazi poizvedovalnega jezika CQL. Pri enostavnem iskanju vtipkamo v iskalno okno besedo in konkordančnik jo bo iskal »na široko«, kar pomeni, da bo iskal bodisi po lemi bodisi po besedni obliki. Napredno iskanje s CQL je sicer bolj zapleteno, vendar omogoča izražanje bistveno kompleksnejših iskalnih pogojev. Uporabljamo lahko tako regularne izraze (npr. [^aeiou]+ za vse pojavnice, ki ne vsebujejo samoglasnikov) ali pa medsebojno kombiniramo iskanje po več pojavnicah, znotraj njih pa po različnih oznakah (npr. [msd-sl="P.*"] [lemma="konj" & msd-sl =".*i"] za nize, ki so sestavljeni iz pridevnika in pojavnic z lemo *konj*, ki so v imenovalniku). Iskanje lahko dodatno omejimo, npr. tako, da mora biti celotno zaporedje v enem stavku ali pa, da iskanje poteka le znotraj določenega dela korpusa, npr. samo v besedilih izbranega avtorja.

Konkordance, torej najdene pojavnice skupaj s kontekstom, lahko izpišemo v centriranem formatu KWIC (*key-word in context*) ali pa kot navadno besedilo

[21] http://www.sketchengine.co.uk/

[22] http://nlp.fi.muni.cz/trac/noske/

z izpostavljenim zadetkom. Pri vsakem zadetku se izpišejo tudi njegovi osnovni metapodatki, če nanje kliknemo, pa vsi. S klikom na posamezen zadetek dobimo njegov širši kontekst, pri CUWI in korpusih IMP pa tudi faksimile ustrezne strani.

Drugi glavni način prikaza rezultatov so frekvenčni slovarji, ki vrnejo samo iskani izraz skupaj s številom pojavitev v korpusu. Frekvenčne slovarje in konkordance je možno razvrstiti na več načinov, pri konkordancah pa lahko tudi izberemo, da nam konkordančnik vrne naključni nabor zadetkov. Pri vseh načinih izpisa velja, da lahko izpišemo ne samo pojavnice (besedilo), temveč katerokoli kombinacijo oznak.

Poleg konkordanc in frekvenčnih slovarjev ponuja noSketchEngine tudi izračun kolokacij za določeno besedo ali besedno zvezo, in to po več različnih statističnih formulah, kot tudi izdelavo podkorpusov in primerjavo besedišča med dvema korpusoma.

Rezultate iskanja lahko shranimo na svoj računalnik in jih tam obdelujemo naprej, pri čemer oba konkordančnika ponujata več formatov za shranjevanje. Oba konkordančnika sta odprta v smislu, da ponujata »govoreče« URL-je, pri katerih lahko kar preko spletnega naslova pridemo do določene poizvedbe; to omogoča citiranje neke poizvedbe kar prek njenega URL-ja ter zajem konkordanc preko programskih vmesnikov.

3.2 Večjezičnost

Večjezičnost v korpusih in konkordančnikih se izraža na več načinov. Oba konkordančnika lahko prikazujeta korpuse v poljubnem jeziku, saj je zapis znakov, tudi interno, v Unikodu. Konkordančnika omogočata tudi prikaz vzporednih poravnanih korpusov, kljub temu da sta CWB in Manatee v zasnovi enojezična. Povezava s poravnanim korpusom je vzpostavljena na ravni strukturnih oznak, ki označujejo poravnane prevodne enote. V konkordančnikih tako vedno iščemo primarno po enem korpusu, vendar lahko prikazujemo tudi poravnane segmente, pri čemer je iskanje mogoče omejiti s postavitvijo dodatnih iskalnih zahtev. Tako lahko npr. iščemo vse pojavitve

leme *predsednik*, pri katerih se v poravnanih angleških segmentih ne pojavi lema *president*.

Večjezičnost je prisotna tudi v oznakah korpusa. Za imena strukturnih in pozicijskih atributov sicer zaenkrat uporabljamo samo angleška imena, imajo pa korpusi zato tipično po dve oznaki za oblikoskladenjsko označitev besede, eno za angleško ime in drugo za slovensko, npr. Ncmsn in Somei za obče samostalnike moškega spola v imenovalniku ednine.

Vsak korpus oz. besedilo ima pripisane tudi metapodatke, ki so v izvornem formatu TEI tipično zapisani tako v slovenskem kot angleškem jeziku. Iz kolofona TEI za vsak korpus oz. besedilo generiramo kolofon v HTML, ki je nato postavljen na splet v dveh različicah – ena ima angleška, druga pa slovenska imena polj in njihovih vrednosti (Erjavec 2010). V obstoječih korpusih zaenkrat na konkordančnikih podpiramo zgolj povezavo do kolofona v slovenskem jeziku.

Zadnji vidik večjezičnosti je jezik, v katerem je napisan vmesnik konkordančnika, oba podpirata tako angleški kot slovenski vmesnik.

5 ZAKLJUČKI

V prispevku smo predstavili korpuse na strežniku nl.ijs.si in konkordančnika, prek katerih je mogoče do njih dostopati. Prispevek je v prvi vrsti namenjen evidentiranju korpusov, ki v marsičem nadgrajujejo dosedanjo ponudbo jezikovnih virov slovenskega jezika, manj pa različnim vprašanjem, ki se pojavijo pri zagotavljanju prostega in trajnega dostopa do takšne količine podatkov, o čemer le na kratko v nadaljevanju.

Problem, ki tradicionalno spremlja prosti dostop do korpusnih podatkov, so avtorske pravice za izvirna besedila in varovanje osebnih podatkov. Pri nekaterih predstavljenih korpusih to ni problematično, npr. pri korpusih starejše slovenščine, pri katerih smo za digitalne predloge, ki jih je zagotovil NUK, sklenili dogovor o prostem razširjanju, ali za korpuse SSJ, pri katerih so bile z besedilodajalci sklenjene ustrezne pogodbe. Za druge korpuse, npr. korpus KONJI, je bil dosežen ustni sporazum o objavi besedil v sklopu

konkordančnika, zataknilo pa se je pri podpisu pisnega dovoljenja. Vendar so besedila v vseh primerih pospremljena z navedbo vira, prek konkordančnikov pa so dostopni samo iztržki, ob tem pa je dostop tudi brezplačen. Besedila, zbrana s spleta (trenutno jpWaC), seveda nimajo pripadajočih dovoljenj, vendar so ta tako ali tako v celoti javno dostopna, ob vsakem pa je naveden tudi njegov URL.

Bolj občutljiva sta korpusa Tweet-sl in ELIZA, vendar ne toliko zaradi avtorskih pravic, pač pa zaradi varovanja osebnih podatkov. V korpus Tweet-sl imen pošiljateljev nismo vključili, hkrati pa so tviti že izvorno javna besedila. Najbolj sporen je korpus ELIZA, problem pa, kot rečeno, rešujemo z zaklepanjem korpusa za javni dostop.

Osnovno načelo je torej omogočiti čim bolj odprt dostop do besedil (Erjavec 2009), težave pa reševati sproti, ko nastanejo: v primeru upravičenih pritožb nad dostopom do posameznih korpusov oz. besedil lahko posamezna besedila iz korpusa odstranimo ali pa korpus zaklenemo.[23]

Poleg prostega je pomembno zagotoviti tudi stalen dostop do korpusov, saj spletne storitve, ki včasih delujejo, včasih pa ne, oziroma nekega dne ne obstajajo več, niso uporabne za resne študije ali vključitev v pedagoški proces. V evropskem prostoru poteka več projektov, ki se trudijo zagotoviti raziskovalne infrastrukture, ki naj bi omogočile stalen dostop do jezikovnih virov in orodij bodisi za namene razvoja jezikovnih tehnologij (META-SHARE)[24] ali za uporabo v humanističnih, predvsem jezikoslovnih raziskavah (CLARIN[25] in deloma DARIAH[26]). Naš strežnik trenutno ne izpolnjuje pogojev za robustno infrastrukturo, vendar je to naš cilj za prihodnost.

Nadaljnje delo tako po eni strani vključuje zagotovitev bolj robustne,

[23] Že sedaj konkordančnika denimo ne dovolita iskalnikom, kot sta Google in Najdi.si, indeksacijo njune vsebine. S tem denimo preprečimo, da bi poizvedba za določeno lastno ime v Googlu, našla tudi pripadajoče konkordance v korpusih.

[24] www.meta-net.eu/meta-share

[25] www.clarin.eu

[26] www.dariah.eu

redundantne računalniške platforme, po drugi pa nadaljevanje dela pri korpusih in vmesnikih. Kljub razmeroma velikemu številu korpusov, ki so že vključeni v konkordančnika, jih ostaja še nekaj, ki bi jih v bodoče radi dodali; v mislih imamo predvsem spletna korpusa hrvaškega in slovenskega jezika hrWaC in slWaC (Ljubešić in Erjavec 2011).

Tudi zapis jezikoslovnih oznak, po katerih lahko iščemo prek konkordančnika, bi bilo mogoče izboljšati. Oznake JOS, IMP in SPOOK je namreč mogoče izraziti tudi kot pare *lastnost* = *vrednost*. Če bi take oblikoskladenjske lastnosti dodali kot pozicijske lastnosti konkordančnikom, bi bilo mogoče enostavno iskati tudi po njih, npr. s poizvedbami, ki bi vrnile besede v orodniku ne glede na besedno vrsto ali naslonke.

Kot je bilo že omenjeno, so korpusi in oblikoskladenjske oznake dostopni prek vmesnika v angleškem in slovenskem jeziku. Polno dvojezičnost, ki je sicer naš cilj, pa je težko vzdrževati tako v pozicijskih kot strukturnih (metapodatkovnih) oznakah. Rešitev, ki bi jo radi implementirali v prihodnosti, je možnost dostopa do vseh korpusov v dveh različicah, od katerih je ena usmerjena v angleški, druga pa v slovenski jezik.

ZAHVALA

Avtor se zahvaljuje vsem, ki so prispevali korpuse, opisane v prispevku, posebej pa Janu Joni Javoršku za implementacijo konkordančnika CUWI in Nikoli Ljubešiću za instalacijo konkordančnika noSketchEngine. Zahvala gre tudi Darji Fišer, anonimnima recenzentoma in urednicama za koristne pripombe na vsebino ter obliko prve različice tega prispevka. Delo na konkordančnikih sta omogočila projekt *ARRS J6-2009-0581 Slovensko prevodoslovje – viri in raziskave* ter program *ARRS P2-0103 (B) Tehnologije znanja*.

LITERATURA

Arhar, Š. (2009): Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo*, 54 (3–4): 43–56.

Arhar Holdt, Š., in Gorjanc, V. (2007): Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52 (2): 95–110.

Christ, O. (1994): A Modular and Flexible Architecture for an Integrated Corpus Query System. *Proceedings of the Conference in Computational Lexicography, COMPLEX '94*: 23–32. Budimpešta: Hungarian Academy of Sciences.

Erjavec, T. (2002): The IJS-ELAN Slovene-English parallel corpus. *International Journal of Corpus Linguistics*, 7 (1): 1–20.

Erjavec, T., Ignat, C., Pouliquen, B., in Steinberger, R. (2005): Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. *Proceedings of the 2nd Language & Technology Conference*: 32–36. Poznan.

Erjavec, T., in Krek, S. (2008): Oblikoskladenjska priporočila in označeni korpusi JOS. *Zbornik Šeste konference Jezikovne tehnologije*: 49–53. Ljubljana: Institut »Jožef Stefan«.

Erjavec, T. (2009): Odprtost jezikovnih virov za slovenščino. *Infrastruktura slovenščine in slovenistike (28. simpozij Obdobja)*: 115–121. Ljubljana: Znanstvena založba Filozofske fakultete.

Erjavec, T. (2010): Text Encoding Initiative Guidelines and their Localisation. *Infoteka*, 11 (1): 3a–14a.

Erjavec, T. (2011): Automatic Linguistic Annotation of Historical Language: ToTrTaLe and XIX Century Slovene. *Proceedings of the 5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH* 2011: 33–38. Portland: Association for Computational Linguistics.

Erjavec, T. (2012): Jezikovni viri starejše slovenščine IMP: zbirka besedil, korpus, slovar. *Zbornik Osme konference Jezikovne tehnologije*: 52–56. Ljubljana: Institut »Jožef Stefan«.

Erjavec, T. (2013): Vzporedni korpus SPOOK: označevanje, zapis in iskanje. V

Š. Vintar. (ur.): *Slovenski prevodi skozi korpusno prizmo*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Erjavec, T., Hmeljak Sangawa, K., Srdanović, I., in Vahčič, A. (2004): Making an XML-Based Japanese-Slovene Learners' Dictionary. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*: 1059–1062. Pariz: European Language Resources Association.

Erjavec, T., Fišer, D. Krek, K., in Ledinek, N. (2010): Jezikovni viri projekta JOS. *Zbornik Sedme konference Jezikovne tehnologije*, 42–48. Ljubljana: Institut »Jožef Stefan«.

Grčar, M., Krek, S., in Dobrovoljc, K. (2012): Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Zbornik Osme konference Jezikovne tehnologije*, 89–94. Ljubljana: Institut »Jožef Stefan«.

Hmeljak Sangawa, K., Erjavec, T., in Kawamura, Y. (2010): Automated Collection of Japanese Word Usage Examples from a Parallel and a Monolingual Corpus. *Proceedings of eLex »eLexicography in the 21st Century: New Challenges, New Applications«*: 137–147. Louvain: Presses Universitaires de Louvain.

Hmeljak Sangawa, K., in Erjavec, T. (2008): A Low Cost Approach to Building a Japanese-Slovene Parallel Corpus. *Denshi Jäohäo Tsäushin Gakkai gijutsu kenkyäu häokoku*, 108: 7–10.

Holozan, P., in dr. (2008): Projekt »Sporazumevanje v slovenskem jeziku«: Specifikacije za učni korpus. Dostopno prek: http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik2/SSJ_Kazalnik_2_Specifikacije-ucni-korpus_v1.pdf.

Jakopin, P., in Michelizza, P. (2007): Besedilni korpus Nova beseda. *Mostovi*, 41 (1/2): 165–176.

Košir, M. (2010). Slovenska filmska terminologija v korpusu filmskih kritik:

Magistrsko delo. Nova Gorica.

Ljubešić, N., in Erjavec, T. (2011): hrWac and slWac: Compiling Web Corpora for Croatian and Slovene. *Lecture Notes in Computer Science 9743*: 395–402. Springer.

Logar Berginc, N., in dr. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Mezeg, A. (2010): Compiling and Using a French-Slovenian Parallel Corpus. *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies* (UCCTS 2010). Ormskirk: Edge Hill University.

Mezeg, A. (2011): Korpusno podprta analiza francoskih polstavkov in njihovih prevedkov v slovenščini: Doktorska disertacija. Ljubljana.

Orwell, G. (1949): *Nineteen Eighty-Four: A novel*. London: Secker & Warburg.

Orwell, G. (1984): *1984*. Ljubljana: Mladinska knjiga (prevod: Alenka Puhar).

Plahuta, H. (2010): Korpusne metode v jezikoslovju pri izdelavi osnutka konjeniškega terminološkega slovarja: Magistrsko delo. Nova Gorica.

Puc, K., in Erjavec T. (2006): Uporaba korpusa pri urejanju spletnega terminološkega slovarja. *Zbornik Pete konference Jezikovne tehnologije*: 156–161. Ljubljana: Institut »Jožef Stefan«.

Rychlý, P. (2007): Manatee/Bonito – A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*: 65–70. Brno: Masaryk University.

Shaw, G. B. (1916): *Pygmalion*. New York: Brentano.

Shaw, G. B. (1997): *Pigmalion*. Celje: Slovensko ljudsko gledališče Celje (prevod: Janko Moder).

Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision

Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester.

Smailović, J., in Pollak, S. (2011): Semi-Automated Construction of a Topic Ontology from Research Papers in the Domain of Language Technologies. *LTC'11, 5th Language & Technology Conference*: 121–125. Poznań.

Steinberger, R., Eisele A., Klocek, S., Pilos, S., in Schlüter, P. (2012): DGT-TM: A Freely Available Translation Memory in 22 Languages. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*: 454–459. Pariz: European Language Resources Association.

Srdanović, I., Erjavec, T., in Kilgarriff, A. (2008): A Web Corpus and Word Sketches for Japanese. *Shizen gengo shori*, 15 (2): 137–159.

Tancig, P., in Žagar, I. (1989): Računalniško podprta analiza velikih tekstualnih baz podatkov: Primer napadov na JNA. *Zbornik V. kongresa Zveze društev za uporabno jezikoslovje Jugoslavije*: 51–56. Ljubljana.

TEI Consortium, ur. (2011): *TEI P5: Guidelines for Electronic Text Encoding and Interchange: Version 1.9.1*. Dostopno prek: http://www.tei-c.org/Guidelines/P5/.

Verdonik, D., in Zwitter Vitez, A. (2011): *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Vintar, Š. (ur.) (2013): *Slovenski prevodi skozi korpusno prizmo*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Vintar, Š., in Erjavec, T. (2000): In Two Minds: How to Teach Translation Students to Learn from Parallel Corpora. *Proceedings of the 4th International Conference on Teaching and Language Corpora*, TaLC 2000: 65. Gradec.

Weizenbaum, J. (1966): ELIZA – A Computer Program for the Study of

Natural Language Communication between Man and Machine. *Communications of the ACM*, 9/36. MIT.

Željko, M., in Krstič, A. (2002): Evroterm – terminologija EU na internetu. *Zbornik devetega posvetovanja Dnevi slovenske* informatike: 408–412. Ljubljana: Slovensko društvo Informatika.

CORPORA AND CONCORDANCERS ON THE NL.IJS.SI SERVER

The paper presents the monolingual and parallel corpora which can be accessed through two concordancers on the server nl.ijs.si. Twelve monolingual corpora contain Slovene language texts, one contains Japanese and one English texts, and comprise reference corpora, such as Gigafida for written contemporary Slovene, IMP for historical Slovene, and GOS for spoken Slovene and specialised corpora, such as the corpus of texts from the informatics domain and the corpus of Slovene tweets. The five parallel corpora contain Slovene texts sentence aligned with, variously, English, Japanese, French, German, and Italian from domains such as EU law, literature and journalism. Although most of the corpora have been produced in the past, they have now been newly annotated, some have been extended with additional texts, and a few are completely new. The texts in the corpora are supplied with meta-data, while their word tokens are either manually or automatically annotated with at least lemmas and morphosyntactic descriptions. Most of the corpora are freely available through two web concordancers, the noSketch Engine and CUWI. These two corpus analysis tools support searching large annotated corpora, various types of search result display, the possibility to filter the searches according to meta-data, and saving the search results locally. In addition to the corpora and concordancers the paper also discusses some issues pertaining to such a corpus-linguistic infrastructure, and concludes with directions for further work.

Keywords: language corpora, concordancers, CWB, CUWI, noSketchEngine