Vzporedni korpus SPOOK: označevanje, zapis in iskanje terminoloških virov

Tomaž Erjavec
Odsek za tehnologije znanja, Institut »Jožef Stefan«
tomaz.erjavec@ijs.si

Abstract
The paper discusses the technical aspects of the compilation and use of the SPOOK parallel corpus, in particular the morphosyntactic tagging and lemmatisation, the encoding of the corpus, and the concordancer developed for searching the corpus. The morphosyntactic tagging of the Slovene texts was performed using previously developed methods and the tagset elaborated in the JOS and MULTEXT-East projects. For tagging the English, French, Italian and German parts of the corpus we used the program TreeTagger, which, however, uses incompatible tagsets for the languages. For SPOOK we developed mappings between the TreeTagger tagsets and the multilingual MULTEXT(-East) recommendations, thus harmonising the tagsets across all the included languages, which enables easier use of the corpus. The computer encoding of the corpus is based on the XML standard. In the process of the compilation of the corpus we used a simple set of Slovene language elements, while the final version follows the Text Encoding Initiative Guidelines, TEI P5. The Web-based concordancer CUWI was also developed for SPOOK, which supports searching the text and annotations through regular expressions, displaying of concordances and frequency lexicons, multilingual searching and display, and downloading results in several formats.

Keywords: morphosyntactic tagging, corpus encoding, concordancer

Ključne besede: oblikoskladenjsko označevanje, kodiranje korpusov, konkordančnik

1 UVOD

Izdelava vzporednega korpusa ima več korakov, začenši z digitalizacijo besedil, poravnavo izvornega besedila in prevoda po stavkih do zapisa metapodatkov za posamezna besedila – v prispevku teh korakov ne obravnavamo, temveč se osredotočimo na jezikoslovno označevanje ter na računalniški zapis korpusa. Te informacije niso zanimive samo s stališča priprave korpusa, temveč tudi s stališča njegove uporabe, saj so jezikoslovne lastnosti, pripisane posameznim besedam v korpusu, lahko koristne pri iskanju po korpusu ter pri luščenju raznih kvantitativnih lastnosti besedil. Korpus, označen z metapodatki o posameznih besedilih in jezikoslovnimi lastnosti posameznih besed, je dostopen preko konkordančnika, ki je tudi predstavljen v prispevku. Konkordančnik je bil razvit za namene prevodoslovnih raziskav in omogoča raznovrstna iskanja po jezikoslovno označenih vzporednih korpusih.

2 OBLIKOSKLADENJSKO OZNAČEVANJE IN LEMATIZACIJA

2.1 Postopek označevanja

Oblikoskladenjsko označevanje je postopek, pri kateri vsaki besedni pojavnici v besedilu glede na sobesedilo pripišemo ustrezno oblikoskladenjsko oznako. Avtomatsko označevanje tipično poteka s statistično zasnovanimi in jezikovno neodvisnimi programi, ki se modela jezika naučijo na ročno označenem korpusu. Modeli so večinoma sestavljeni iz dveh delov:

- Leksikon besednim oblikam pripiše množice oblikoskladenjskih oznak, npr. »*čebule* → *Sozer:8* + *Sozmt:1*, torej 8x samostalnik, občno_ime, ženski spol, ednina, rodilnik in 1x v tožilniku množine.
- Kontekstne značilke, s pomočjo katerih iz množice leksikalnih oznak program izbere pravo glede na kontekst besede. Tipične značilke so kar n-terčki oznak s pripadajočo frekvenco iz učnega korpusa, npr. *Zk-mer Ppnmer Somer:46*, primer niza je »*tega mestnega orkestra*« ali »*onega narodnega veljaka*«

Za označevanje slovenskega dela korpusa smo uporabili orodje ToTaLe (Erjavec et al., 2005), ki besedilo najprej tokenizira, torej razdeli na pojavnice (besede in ločila), nato tagira (oblikoskladenjsko označi) in na koncu lematizira (besednim pojavnicam pripiše osnovno obliko). Program se je modela jezika (leksikona in kontekstnih značilk označevalnika, kot tudi modela lematizacije) naučil na korpusu jos1M (Erjavec in Krek 2008), ki vsebuje milijon besednih pojavnic, označenih z oblikoskladenjsko oznako in lemo. Program je uporabil tudi pomožni

leksikon, izluščen iz korpusa FidaPLUS (Arhar in Gorjanc, 2007) – ta je sicer zelo velik, vendar vsebuje napake avtomatskega označevanja in je zato manj zanesljiv kot leksikon iz jos1M. Nad korpusom SPOOK nismo naredili evalvacije točnosti označevanja, je pa bila ta narejena v okviru projekta JOS (Erjavec in Krek 2008), kjer je oblikoskladenjsko označevanje doseglo točnost 86,6 %, pri čemer je bila evalvacija narejena z učenjem in desetkratnim križnim preverjanjem neposredno na korpusu jos100k. Ker je po eni strani učni korpus in pomožni leksikon modela za označevanje SPOOK bistveno večji, kot je bil pri evalvaciji točnosti z jos100k, po drugi pa besedilni tip besedil drugačen kot v jos100k, ni jasno, koliko se točnost označevanja slovenskih besedil v SPOOK razlikuje od te vrednosti.

Za označevanje angleškega, francoskega, italijanskega in nemškega dela korpusa smo uporabili označevalnik TreeTagger (Schmid 1994; TreeTagger – a language independent part-of-speech tagger), ki je prosto dostopen in ima prednost, da je na voljo skupaj z modeli za vse te jezike. To zelo olajša označevanje, saj se izognemo dolgotrajnemu in kompleksnemu postopku zagotovitve ročno označenih korpusov za vsakega od jezikov in šolanju označevalnika. Tudi za TreeTagger nismo naredili evalvacije točnosti označevanja, ta pa tudi sicer, kolikor nam je znano, ne obstaja za vse jezike, z izjemo angleščine (Schmid 1994), kjer je ocenjena na 96,3 %, in nemščine (Schmid 1995), kjer je ocenjena na 97,5 %. Te številke so sicer mnogo boljše kot za slovenski jezik, vendar so tudi nabori oblikoskladenjskih oznak bistveno manjši kot za slovenščino. Če ima slovenščina v sistemu JOS preko 1,900 različnih oznak, jih ima od drugih jezikov v SPOOK največ nemščina (52), najmanj pa francoščina (30). Razlogi za razlike so deloma jezikoslovni, saj so slovanski jeziki bolj pregibni od zahodnoevropskih, izvirajo pa tudi iz pragmatičnih ali z lokalno jezikoslovno tradicijo povezanih odločitev avtorjev posameznih naborov.

Tako ToTaLe kot TreeTagger poleg oblikoskladenjskega označevanja besedilo tudi lematizirata, vsaki besedni obliki v besedilu torej pripišeta njeno osnovno obliko. Lematizacija za slovenščino prav tako temelji na modelu, naučenem na (leksikonu) jos1M, pri čemer je program sposoben razmeroma natančno lematizirati tudi neznane besede. Za razliko od ToTaLe TreeTagger podpira samo lematizacijo znanih besed, vendar pa za lematizacijo zahodnoevropskih jezikov že to ne daje slabih rezultatov, saj so bili učeni na velikih korpusih, in gre za jezike, ki so za razliko od slovenščine pregibno precej bolj enostavni.

2.2 Oblikoskladenjska priporočila SPOOK

Modeli TreeTagger za tuje jezike korpusa SPOOK so bili naučeni iz večjih dostopnih ročno označenih korpusov za posamezne jezike, vendar pa ti korpusi uporabljajo med seboj neodvisne nabore oznak. Tako je npr. oznaka za lastno ime

za angleščino *NP*, za nemščino *NE*, za francoščino *NAM* in za italijanščino *NPR*. Takšno neskladje seveda otežuje uporabo korpusa kot celote, zato smo poskusili uskladiti nabore oznak (in oblikoskladenjskih lastnosti) za vse jezike korpusa SPOOK, tako da smo naredili preslikave izvornih TreeTagger (in ToTaLe) oznak v analitični sistem narejen po metodologiji MULTEXT.

2.2.1 Oblikoskladenjska priporočila MULTEXT-East in JOS

Za slovenski jezik uporabljamo priporočila za oblikoslovno označevanje JOS (Erjavec in Krek 2008; Oblikoskladenjske specifikacije JOS V1.1), oz. kar ustreza slovenskemu delu večjezičnih oblikoskladenjskih priporočil MULTEXT-East (Erjavec (ur.) 2010a; Erjavec 2012a). Osnovni princip teh priporočil je, da za vsak jezik definirajo nabor oblikoskladenjskih oznak (npr. ena od oznak za slovenščino je *Ncfsg*), definirajo pa tudi sisteme lastnosti in enostavno preslikavo oznak v te sisteme, npr. *Ncfsg* ≡ *Noun Type=common Gender=feminine Number=singular Case=genitive.* Priporočila podpirajo tudi prevajanje oz. lokalizacijo oznak in lastnosti v druge jezike, npr. *Ncfsg* ≡ *Sozer* ≡*samostalnik vrsta=občno_ime spol=ženski število=ednina sklon=rodilnik.*

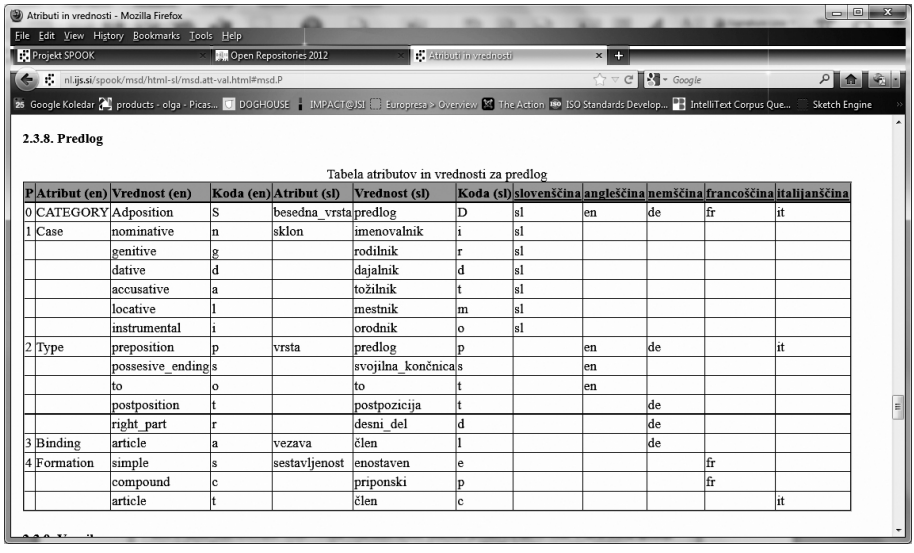
Priporočila MULTEXT-East pokrivajo 16 jezikov, med njimi veliko večino slovanskih, ne vsebujejo pa zahodnoevropskih jezikov. Oblikoskladenjski nabori oznak za te so bili sicer definirani v okviru prvotnega projekta MULTEXT (Ide in Veronis 1994), vendar bi težko dobili ročno označene korpuse za vse jezike SPOOK, ki bi uporabljali nabore MULTEXT.

Da bi uskladili nabore oznak, smo po vzoru priporočil MULTEXT-East naredili petjezična priporočila SPOOK, ki so v slovenskem delu (skoraj) enaka kot JOS, v priporočilih za tujejezično označevanja pa uporabljajo metodologijo večjezičnih definicij MULTEXT(-East), pri čemer imajo njihove oznake enostavno, 1-1 preslikavo z izvornimi oznakami programa TreeTagger.

Priporočila imajo razmeroma enostavno strukturo. Najprej definirajo *kategorije* oz. besedne vrste, pri čemer so te »besedne vrste« ponekod tudi tehnične narave, npr. *okrajšava* ali *neuvrščeno.* Angleška in slovenska imena ter enočrkovne kode za 12 definiranih kategorij MULTEXT podaja tabela 1. V tabeli je tudi število atributov za posamezno kategorijo. Kategorijam namreč priporočila pripišejo nabor lastnosti, tj. atributov in naborov njihovih vrednosti. Te, t. i. skupne tabele, pokrivajo vse jezike in pri vsaki kombinaciji atributa in vrednosti za kategorijo tudi določijo, za katere jezike se uporablja. Primer take tabele, vzet neposredno s spleta, je podan v sliki 1.

Tabela 1: Kategorije MULTEXT-East oz. SPOOK

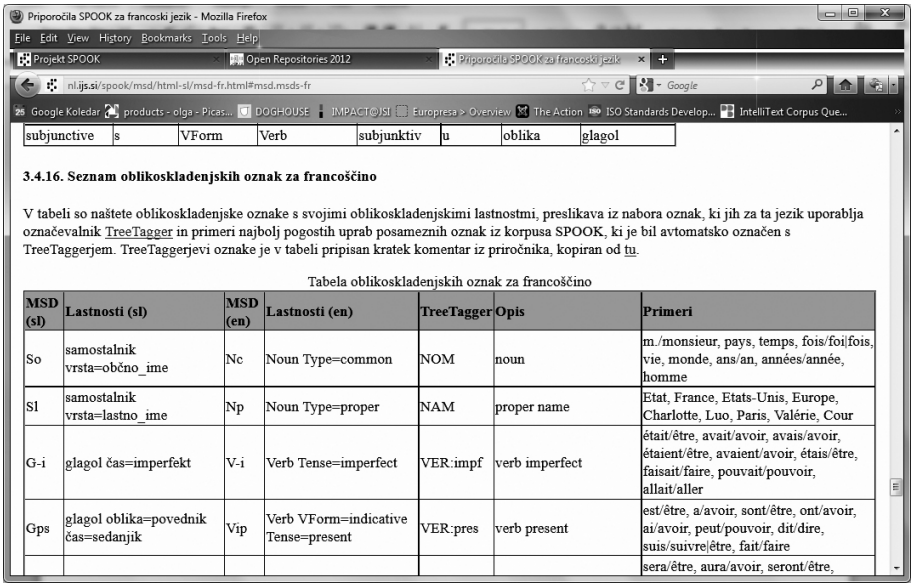
Kategorija (en)	Koda (en)	Kategorija (sl)	Koda (sl)	Atributov	Definirano za jezike				
Noun	N	samostalnik	S	5	sl	en	de	fr	it
Verb	V	glagol	G	10	sl	en	de	fr	it
Adjective	A	pridevnik	P	7	sl	en	de	fr	it
Pronoun	P	zaimek	Z	11	sl	en	de	fr	it
Determiner	D	določilnik	I	1		en			it
Article	T	člen	C	1			de	fr	it
Adverb	R	prislov	R	2	sl	en	de	fr	it
Adposition	S	predlog	D	4	sl	en	de	fr	it
Conjunction	C	veznik	V	2	sl	en	de	fr	it
Numeral	M	številnik	K	6	sl	en	de	fr	it
Particle	Q	členek	L	2	sl	en	de		it
Interjection	I	medmet	M	0	sl	en	de	fr	
Abbreviation	Y	okrajšava	O	1	sl	en	de	fr	
Residual	X	neuvrščeno	N	1	sl	en	de		it



P	Atribut (en)	Vrednost (en)	Koda (en)	Atribut (sl)	Vrednost (sl)	Koda (sl)	slovenščina	angleščina	nemščina	francoščina	italijanščina
0	CATEGORY	Adposition	S	besedna_vrsta	predlog	D	sl	en	de	fr	it
1	Case	nominative	n	sklon	imenovalnik	i	sl				
		genitive	g		rodilnik	r	sl				
		dative	d		dajalnik	d	sl				
		accusative	a		tožilnik	t	sl				
		locative	l		mestnik	m	sl				
		instrumental	i		orodnik	o	sl				
2	Type	preposition	p	vrsta	predlog	p		en	de		it
		possesive_ending	s		svojilna_končnica	s		en			
		to	o		to	t		en			
		postposition	t		postpozicija	t			de		
		right_part	r		desni_del	d			de		
3	Binding	article	a	vezava	člen	l			de		
4	Formation	simple	s	sestavljenost	enostaven	e				fr	
		compound	c		priponski	p				fr	
		article	t		člen	c					it

Slika 1: Definicija za kategorijo »predlog« v skupni tabeli oblikoskladenjskih priporočil

Vsak jezik ima nato v priporočilih še svoj razdelek, kjer so zapisane podobne tabele, kot so skupne, vendar samo za kategorije in lastnosti, ki so za ta jezik definirane. Razdelki za posamezne jezike pa vsebujejo tudi nabore oblikoskladenjskih oznak, ki so na enostaven način sestavljene iz kode kategorije in kod za posamezne vrednosti naštetih atributov – te oznake se, ker so kratke, uporablja za označevanje korpusov, čeprav so formalno enake sistemom lastnosti. Primer seznama definiranih oznak za italijanščino je dan v sliki 2, kjer je vsaka oznaka podana dvojezično, pripisani pa so ji tudi primeri uporabe, pri čemer so ti vzeti kar iz označenega korpusa.



subjunctive	s	VForm	Verb	subjunktiv	u	oblika	glagol

3.4.16. Seznam oblikoskladenjskih oznak za francoščino

V tabeli so naštete oblikoskladenjske oznake s svojimi oblikoskladenjskimi lastnostmi, preslikava iz nabora oznak, ki jih za ta jezik uporablja označevalnik TreeTagger in primeri najbolj pogostih uprab posameznih oznak iz korpusa SPOOK, ki je bil avtomatsko označen s TreeTaggerjem. TreeTaggerjevi oznake je v tabeli pripisan kratek komentar iz priročnika, kopiran od tu.

Tabela oblikoskladenjskih oznak za francoščino

MSD (sl)	Lastnosti (sl)	MSD (en)	Lastnosti (en)	TreeTagger	Opis	Primeri
So	samostalnik vrsta=občno_ime	Nc	Noun Type=common	NOM	noun	m./monsieur, pays, temps, fois/foi|fois, vie, monde, ans/an, années/année, homme
Sl	samostalnik vrsta=lastno_ime	Np	Noun Type=proper	NAM	proper name	Etat, France, Etats-Unis, Europe, Charlotte, Luo, Paris, Valérie, Cour
G-i	glagol čas=imperfekt	V-i	Verb Tense=imperfect	VER:impf	verb imperfect	était/être, avait/avoir, avais/avoir, étaient/être, avaient/avoir, étais/être, faisait/faire, pouvait/pouvoir, allait/aller
Gps	glagol oblika=povednik čas=sedanjik	Vip	Verb VForm=indicative Tense=present	VER:pres	verb present	est/être, a/avoir, sont/être, ont/avoir, ai/avoir, peut/pouvoir, dit/dire, suis/suivre|être, fait/faire
						sera/être, aura/avoir, seront/être,

Slika 2: Začetek tabele oblikoskladenjskih oznak SPOOK za francoščino.

Pri preslikavi oznak TreeTagger v sistem MULTEXT-East največ problemov povzroča razlika med analitičnimi oznakami slednjega in pogosto sintetičnimi oznakami prvega. Tako ima npr. angleščina oznako *VBZ* za *Verb, 3rd person singular present,* kar je preslikano v *G-iste* oz. *glagol vrsta=nedoločeno oblika=povednik čas=sedanjik oseba=tretja število=ednina*. To sicer ni problematično, vendar ima angleščina tudi oznako *VBP* oz. *Verb, non-3rd person singular present*, za katero je dosti manj jasno, kako jo ustrezno preslikati. Ker smo želeli ohraniti bijektivno (1:1) preslikavo med obema sistemoma oznak, smo se v tem primeru odločili, da osebi in številu pustimo nedoločeno vrednost, torej *glagol oblika=povednik čas=sedanjik* torej *G-is--* oz. kar *G-is*.

Problematične so tudi oznake TreeTagger, ki so lahko pripisane posamezni besedi, npr. oznaka *CHE* za italijansko besedico *che*. Take oznake se uporabljajo za zelo

dvoumne funkcijske besede, ki jih oblikoskladenjski označevalniki ne zmorejo razdvoumiti, z besedi lastno oznako pa se temu problemu enostavno izognejo. Za zapis takšnih primerov v priporočilih SPOOK besedi pripišemo (idealno zanjo najbolj prototipično) besedno vrsto, nato pa jo kvalificiramo s posebno lastnostjo. Tako se *CHE* preslika v *členek vrsta=che* oz. *Lh*.

2.2.2 Oblikoskladenjska priporočila IMP

Kot rečeno, so slovenske oblikoskladenjske oznake oz. lastnosti že v sistemu JOS oz. MULTEXT-East. Priporočila SPOOK vsebujejo več ali manj kopijo le-teh, vendar z manjšo razliko. Ker imajo ostali jeziki dva sistema oznak, pri čemer jih ločijo 30–50, smo tudi v slovenska priporočila (in korpus) dodali preslikavo v manjši nabor oznak.

Ta nabor smo izdelali za označevanje korpusov starejšega slovenskega jezika IMP (Erjavec 2012b; Jezikovni viri starejšega slovenskega jezika IMP), kjer je bil poudarek na označevanju besed s posodobljenimi oblikami, vseeno pa smo želeli v korpusu imeti tudi ročno preverjene oznake vsaj za besedno vrsto in manjši nabor leksikalnih lastnosti. Zato smo izdelali priporočila IMP, ki so enaka kot JOS, samo da izpustijo vse pregibne lastnosti, s čimer dobimo nabor 32 oznak, kar se tudi lepo sklada s številom oznak ostalih jezikov.

Preslikava oznak JOS v nabor IMP je zelo enostavna, saj oznake JOS samo okrajšamo, s čimer izgubimo pregibne lastnosti besed, ohranimo pa besedno vrsto in (nekaj) inherentnih lastnosti. Tako ima npr. nabor IMP namesto 279 oznak za pridevnike, kolikor jih ima JOS, samo 5 oznak, kjer je določena vrsta (*splošni, svojilni, deležniški*) in pa, za splošne pridevnike, stopnja (*nedoločeno, primernik, presežnik*). Slovenščina tako v SPOOK-u vsebuje preslikavo med naborom oznak JOS in naborom IMP. Kot drugi jeziki torej uporablja dva nabora oznak, čeprav je njihov status drugačen kot za druge jezike.

2.2.3 Dostopnost priporočil

Priporočila SPOOK so na voljo na enak način kakor MULTEXT-East, JOS in IMP, pri katerih smo se trudili, da bi bila čim bolj dostopna. Dostopnost ima več dimenzij (Erjavec 2009), ki so pri oblikoskladenjskih priporočilih SPOOK udejanjene na sledeče načine:

- za izmenljivost, trajnost in možnost uporabe v raznovrstne namene so kodirana po mednarodnih standardih in priporočilih;

- dostopna so na spletu tako v obliki XML kot v pretvorbi v HTML, torej stavljena s kazalom in navigacijo, v obliki, primerni za branje oz. kot spletna referenca, saj imajo vsa poglavja fiksne URL-je (npr. http://nl.ijs.si/spook/msd/html-sl/msd-de.html msd.P-de kaže na tabelo za nemški zaimek);
- nabori oblikoskladenjskih oznak in lastnosti SPOOK so definirani tako v slovenskem kot v angleškem jeziku, v obeh je napisano tudi tako prosto besedilo priporočil kot tudi stavljenje v HTML (npr. »Kazalo« in »Table of contents«), s čimer je omogočena uporaba slovenščine za naše jezikoslovce, obenem pa razumevanje oznak za tuje raziskovalce;
- avtorskopravno so priporočila dostopna pod licenco »Creative Commons Priznanje avtorstva-Deljenje pod enakimi pogoji« (Creative Commons licence Attribution-ShareAlike 3.0), kar pomeni, da jih je dovoljeno kopirati, distribuirati in prenašati ter tudi spreminjati in uporabljati v komercialne namene pod pogojem, da se navede izvirnega avtorja in, če se jih spreminja, da se jih deli pod enakimi pogoji.

3 RAČUNALNIŠKI ZAPIS IN OBSEG

3.1 Zapis XML

Za računalniški zapis korpusa in oblikoskladenjskih priporočil smo uporabili standard W3C XML (Extensible Markup Language (XML)), ki je prevladujoči način zapisa (tudi) digitalnih besedil. V XML lahko definiramo nabore (besedišča) oznak in dovoljena medsebojna razmerja, te nabore pa potem uporabljamo za označevanje različnih zvrsti besedil. XML ima prednost, da je po eni strani razmeroma berljiv, po drugi pa primeren za strojno obdelavo. Definicije naborov za opis tipov dokumentov so izražene v shemah XML, kar omogoča strojno preverljivost pravilnosti oznak v dokumentih, ki je zelo pomembna pri izdelavi jezikovnih virov, ker zagotavlja, da v dokumentih ni tehničnih napak.

Zapis XML je tudi dobro podprt z orodji, oz. pridruženimi standardi. Eden bolj pomembnih je XSLT (XSL Transformations (XSLT) Version 1.0), ki definira jezik za transformacije med dokumenti XML. V projektu SPOOK smo XSLT skripta uporabili za pretvorbo delovnega formata korpusa v kanoničnega (opis v naslednjem razdelku) in nato za pretvorbo kanoničnega formata v izvedene oblike, primerne za konkordančnik ali statistične analize. Tudi oblikoskladenjska priporočila so zapisana v XML, nato pa z XSLT pretvorjena v HTML za branje, ali v razne tabelarične formate, ki podajo npr. preslikavo oblikoskladenjskih oznak v lastnosti, prevedejo iz slovenščine v angleščino, podajo sortirno zaporedje oznak itd.

V procesu izdelave korpusa smo uporabili dva zapisa v XML. V delovnem formatu je bil zapisan končni rezultat prve faze obdelave korpusnih enot, ki vsebujejo korigirano besedilo, so opremljene z bibliografskimi podatki in so razdeljene na prevodne segmente, ki so poravnani med izvirnikom in prevodom. Format oz. shema XML, ki podpira ta zapis, je bila namensko napisana in uporablja slovenska imena elementov, obenem pa je tudi minimalna po številu uporabljenih elementov in torej primerna za manjše ročne korekcije.

3.2 Nabor znakov Unikod

Standard XML je osnovan na uporabi univerzalnega kodnega nabora Unikod (The Unicode Consortium), ki je tudi uporabljen v korpusu. Vendar pa je korpus v procesu izdelave moral iti preko več programov, ki lahko tako ali drugače pokvarijo kodiranje znakov. Največ problemov je povzročil ravno TreeTagger, saj ta za več jezikov predpostavlja, da bo format besedila v starem, 8-bitnem kodiranju ISO 8896-1 oz. »Latin-1«, ki pokriva zahodnoevropske jezike. Če se v takem besedilu pojavijo ne-zahodnoevropske črke ali posebna ločila, teh ni mogoče zapisati v Latin-1. Zato je bilo treba napisati programe, ki besedilo v XML najprej glede na jezik pretvorijo v format in kodiranje za TreeTagger, nato pa združijo TreeTaggerjev izhod z izvornim besedilom, da ohranijo izvorni in s tem pravilni zapis znakov.

Tudi v delovnem zapisu XML so se še pojavljale napake kodiranja, ki smo jih poskusili popraviti z namenskim programom. Najprej smo naredili profil znakov za celoten korpus in po datotekah, nato pa pogledali, kateri znaki predstavljajo tipične napake, ki jih je mogoče popraviti, in katerih znakov ni mogoče enostavno interpretirati v naboru Unikod – prve smo s programom popravili v ustrezne znake Unikod, slednje pa nadomestili z znakom Unikod U+FFFD z imenom »REPLACEMENT CHARACTER«, ki opozarja, da je bil na tem mestu znak, ki ga v Unikodu nismo mogli predstaviti. Takšnih znakov je malo, v celotnem korpusu jih je samo 114, pojavljajo pa se v 10 dokumentih.

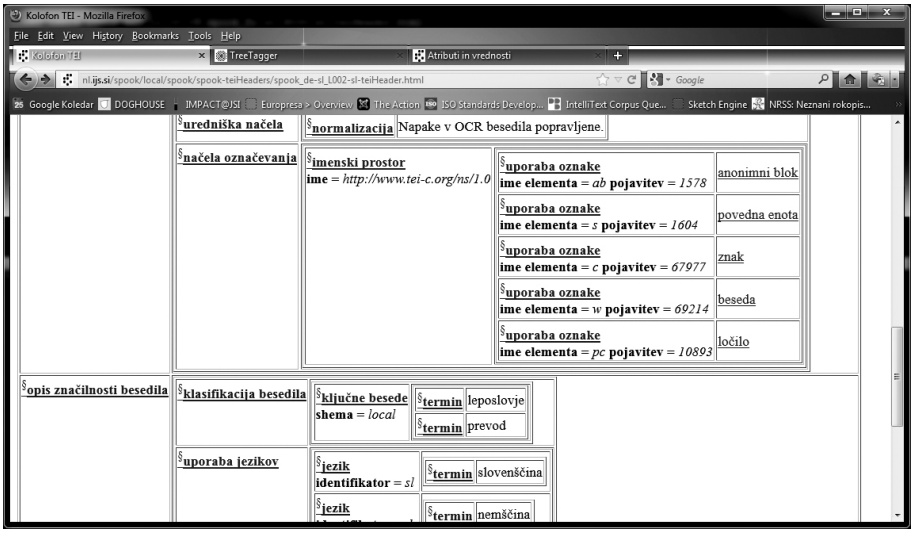
Korpus ima vsega skupaj 52 milijonov znakov besedila vključno s presledki, uporablja pa 270 različnih znakov, tudi npr. črke cirilične in drugih abeced, razne vrste navednic in vezajev ter matematične simbole.

3.3 SPOOK v zapisu TEI

V drugi fazi smo besedila, kot je bilo opisano v prejšnjem poglavju, avtomatsko jezikoslovno označili, obenem pa pretvorili format v takega, ki je skladen s priporočili

za zapis besedil TEI P5 (TEI P5: Guidelines for Electronic Text Encoding and Interchange), torej z de-facto standardom v digitalni humanistiki. Priporočila določajo zapis XML za razne vrste jezikovnega gradiva, ki se uporablja v znanstvene namene, kot so npr. slovarji, tekstnokritične izdaje, opisi rokopisov, itd., pa tudi za jezikoslovno označene korpuse in strukture lastnosti. V projektu SPOOK uporabljamo TEI tako za zapis korpusa kot tudi oblikoskladenjskih priporočil.

V korpusu SPOOK je vsako besedilo zapisano kot dokument TEI, ki se začne s kolofonom, nato pa vsebuje besedilo. TEI v svojem kolofonu ponuja bogat nabor meta-podatkov, od zapisa raznih vrst odgovornosti za besedilo in kodiranje, opis vira, uredniška načela, itd. V sliki 3 je kot primer dan del kolofona enega od besedil – kolofon tu ni neposredno zapisan v TEI/XML, pač pa v spletni obliki, ki je avtomatsko izvedena iz XML in tudi opisno prevede sicer angleška imena elementov v slovenske termine (Erjavec 2010b). Prikazani del kolofona vsebuje kratek opis uredniških načel normalizacije besedila, opis uporabe oznak v besedilu, klasifikacijo besedila kot prevod leposlovnega dela in definicije kod za jezike. Slednje so zanimive s stališča medsebojnega navezovanja standardov: TEI določa, da morajo biti kode jezikov zapisane po standardu ISO 639, s čimer že pravilna uporaba priporočil TEI zagotovi, da bodo tudi kode jezikov (in npr. zapis datumov, URLjev, itd.) sledili standardom in priporočilom, kar bistveno olajša strojno procesiranje takšnih jezikovnih virov.



Slika 3: Del kolofona TEI za eno od besedil korpusa SPOOK.

Strukturo besedila ilustrira slika 4. Besedilo dokumenta je sestavljeno iz prevodnih ustreznic, (*ab*) ki so povezane (preko *@xml:id* in *@corresp*) z istoležnimi ustreznicami iz vzporednega besedila. Prevodne ustreznice so nato sestavljene iz

stavkov oz. povedi (*s*), pri čemer so bili stavki avtomatsko označeni (tako TreeTagger kot ToTaLe naredita vzporedno s tokenizacijo še segmentacijo na stavke), zato segmentacija ni vedno pravilna.

```
<text xml:id="spook_fr-sl_N064-fr.text" xml:lang="fr"
      corresp="spook_fr-sl_N064-sl.xml spook_fr-sl_N064-sl.text">
  <body>
    <ab n="1" xml:id="seg.1" corresp="spook_fr-sl_N064-sl.xml seg.1">
      <s>
        <w lemma="le" ana="T" ctag="DET:ART">Le</w>
        <c> </c>
        <w lemma="takfrisime" ana="Nc" ctag="NOM">takfrisime</w>
        <pc ctag=",">,</pc>
        <c> </c>
        <w lemma="un" ana="T" ctag="DET:ART">une</w>
        <c> </c>
        <w lemma="idéologie" ana="Nc" ctag="NOM">idéologie</w>
        <c> </c>
        <w lemma="messianique" ana="A" ctag="ADJ">messianique</w>
      </s>
    </ab>
```

Slika 4: Poravnano in jezikoslovno označeno besedilo v zapisu TEI.

Stavki so sestavljeni iz besed, ločil in presledkov. Besede (*w*) nosijo tri atribute, in sicer lemo (*@lemma*), oblikoskladenjsko analizo (*@ana* = oznaka SPOOK) in korpusno oznako (*@ctag* = oznaka TreeTagger). Za slovenska besedila pripišemo pregibne oznake JOS h korpusni oznaki, saj je ToTaLe z njimi označil korpus, leksikalne oznake IMP pa oblikoskladenjski analizi, ker so podobne analizam za ostale jezike.

Ločila *(pc*) seveda nimajo oblikoskladenjskih oznak ali lem, so pa označena s korpusno oznako (*@ctag*), pri čemer so te poenotene med jeziki in so kar enake kot samo ločilo.

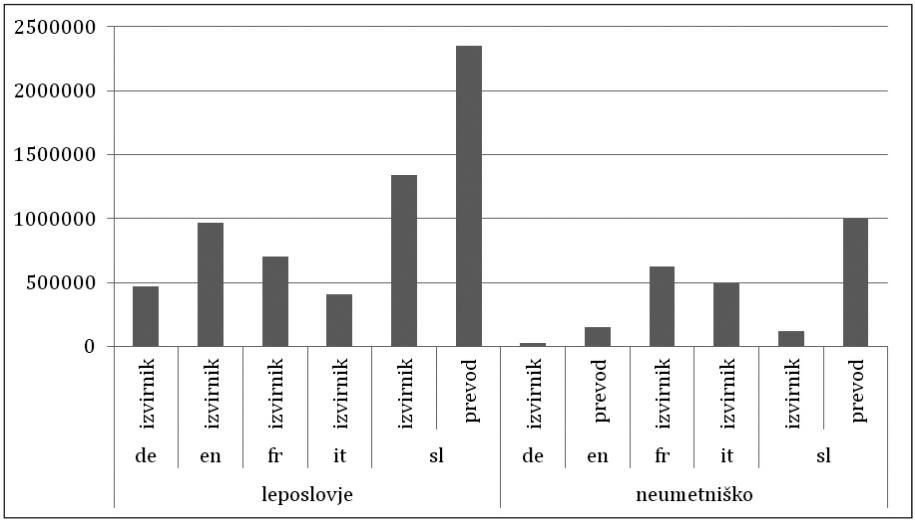
Vključevanje označenih presledkov je v računalniških korpusih prej posebnost kot pravilo, saj so pri jezikoslovnih obravnavah nepomembni. Vendar je, če drugega ne, vsaj lepo, če obdržimo stičnost pojavnic in s tem možnost pravilnega stavljena prevodnih segmentov. Zato so v korpusu SPOOK označeni tudi presledki, in sicer kot presledek v elementu za posamezen znak (*c*, zapisano torej kot <*c*> </*c*>). Element za znak v korpusu SPOOK torej vsebuje vedno in samo presledek.

TEI služi kot kanonični format korpusa, pa tudi oblikoskladenjskih priporočil – ta zapis je dobro dokumentiran, izmenljiv in odporen na tehnološke spremembe. Za uporabo v aplikacijah je format v XML najprej potrebno transformirati v ciljni zapis, kar dosežemo s skriptami XSLT. Primer uporabe teh skript smo že videli na primeru slik spletne predstavitve oblikoskladenjskih priporočil in kolofona TEI. Podoben postopek uporabimo tudi pri konverziji v podatkovno bazo za konkordančnik; več o tem v naslednjem poglavju.

3.4 Velikost korpusa

Ko so besedila označena z metapodatki (npr. uporabljene oznake in njihovo število), ni težko prikazati medsebojnih razmerij med posameznimi jeziki oz. zvrstmi besedil. V tabeli 2 so podane porazdelitve nekaterih metapodatkov iz korpusa SPOOK, kjer so razmerja izražena s številom besed. Kot vidimo, ima SPOOK precej več leposlovnih kot neumetniških besedil, pri čemer so bila tu za namene projekta zbrana besedila s tujejezičnim izvirnikom in prevodom v slovenščino. Izvirna slovenska leposlovna besedila niso bila posebej zbrana za projekt SPOOK, temveč smo vključili romane iz korpusa Gigafida.

Tabela 2: Razmerja med zvrstmi besedil v korpusu glede na število besed.



Kar se velikosti datotek tiče, je posledica berljivosti XML in univerzalnosti TEI precejšnja potratnost s prostorom – celoten označeni korpus, ki vsebuje 8 milijonov besed, je velik več kot 800 MB. Vendar je prostor na disku poceni, datoteke XML pa se tudi učinkovito komprimirajo; stisnjen v formatu ZIP ima SPOOK tako samo 70MB.

4 KONKORDANČNIK CUWI

Za korpus SPOOK je bil razvit konkordančnik po imenu CUWI (Corpus Users' Web Interface), ki omogoča iskanje po besedilih v korpusih in raznovrstne prikaze rezultatov. Konkordančnik podpira iskanje tako po besedah oz. besednih zvezah kot po jezikoslovnih oznakah, podpira pa tudi omejitev iskanj glede na metapodatke besedil.

4.1 Implementacija konkordančnika

Konkordančnik je sestavljen iz zalednega dela, t.j. programske opreme, ki dejansko izvaja iskanje po korpusu, in čelnega dela sistema, torej spletnega vmesnika. Kot zaledni del uporabljamo CWB (Christ 1994; CWB: The IMS Open Corpus Workbench), ki je visokozmogljiv sistem v uporabi za številne korpuse tako pri nas kot po svetu. Za zasnovo iskanja v CWB je ilustrativno pogledati format datotek, ki jih sistem uporablja za posamezen korpus. En korpus je v CWB ena datoteka, ki jo iz kanoničnega formata TEI avtomatsko pretvorimo s skripto XSLT.

Datoteka CWB vsebuje strukturne elemente in označene pojavnice. Strukturni elementi so v resnici kar elementi XML z atributi, pojavnice pa so besede ali ločila, zapisane skupaj z oznakami v tabelaričnem formatu, kot ilustrira slika 5.

```
<text id="FSL001" lang="fr" title="Fou de Vincent"
    title-sl="Nor na Vincenta" author="Hervé Guibert"
    translator="Brane Mozetič" type="leposlovje">
<seg id="FSL001.1" corresp="spook_fraslv-slv.cqp FSL001.1">
<s>
Dans        dans        dans        Ss    De    PRP
la          la          le          T     T     DET:ART
nuit        nuit        nuit        Nc    So    NOM
du          du          du          Sc    Dp    PRP:det
25          25          25          M     K     NUM
au          au          au          Sc    Dp    PRP:det
26          26          26          M     K     NUM
novembre    novembre    novembre    Nc    So    NOM
,           ,           ,           ,     ,     PUN
```

Slika 5: Format zapisa CWB.

CWB omogoča izpis metapodatkov (atributov elementa *text*) posameznih zadetkov v korpusu, pa tudi omejitev iskanja na tista besedila, kjer so metapodatki enaki kot zahtevana omejitev (npr. iskanje samo po prevedenih besedilih). Kot vidimo, je CWB korpus v zasnovi enojezičen, povezava z vzporednim korpusom pa je vzpostavljena na ravni segmentov. Tako v CWB vedno iščemo primarno po enem korpusu, lahko pa prikažemo tudi poravnani segment. Pojavnice in njihove oznake imajo v CWB enakovreden status, tako da je mogoče tako iskanje kot prikaz pojavnic oz. njihovih oznak, ali pa poljubne kombinacije obojih. Kar se tiče formata izpisa, velja še opomba, da CWB ne ohrani stičnosti pojavnic.

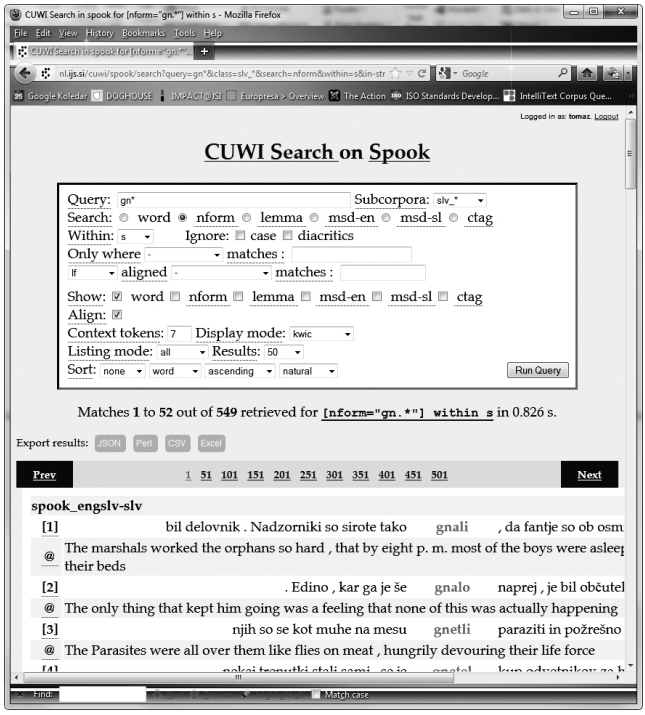
Spletni vmesnik CUWI smo razvili posebej za korpus SPOOK. CUWI je zasnovan na platformi Mojolicious (Mojolicious – Perl real-time web framework; McDaniel 2011, poglavje »Websockets«) za gradnjo spletnih aplikacij in je sestavljen iz zalednega dela CWB::Model, ki omogoča programski dostop (API) do korpusov v formatu CWB, in spletnega vmesnika CWB::CUWI, ki omogoča spletno poizvedovanje in prikaz. Platforma Mojolicious je napisana za jezik Perl, ki je tudi standardni skriptni vmesnik za korpuse CWB, zasnovana pa je po vzorcu reaktorskega demultiplekserja dogodkov (*reactor demultiplexing event loop*, Schmidt 1995), tako da lahko zelo majhno število spletnih servisov hkrati opravlja poizvedbe za več uporabnikov in predstavi rezultate, ko jih zaledni del CWB pripravi. Ta arhitektura omogoča veliko mero fleksibilnosti in zelo dober izkoristek sodobnih računalniških arhitektur z več procesorskimi jedri. Ker so korpusi SPOOK relativno kompleksni, je tako nastal konkordančnik, ki je splošno uporaben in ga je zelo lahko prilagoditi tudi za uporabo z drugimi korpusi in v drugačnih jezikovnih okoljih. Fleksibilna platforma, zgrajena po sodobnem modelu *Model-View-Controller* (MVC, Greer 2007; Burbeck 1992) omogoča tudi preprosto prilagajanje z dodatnimi predlogami za spletne strani (»templates«) in programskim vtičniki (»plugins«). Zaledni del CWB::Model prinaša dodatno možnost, da lahko uporabnik pripravi poizvedbe, ki bodo tekle samodejno, brez spletnega vmesnika, in jih z njegovo pomočjo strojno obdela z lastnim programom. CUWI je prosto dostopen pod dvojno licenco *Perl Artistic Licence / GPL2.*

4.2 Uporaba konkordančnika

CUWI podpira raznovrstna iskanja in prikaze rezultatov iz korpusa SPOOK skozi masko, ki je podana v sliki 6**.** Iskanje je lahko enostavno ali pa z izrazi CWB. Pri enostavnem iskanju vtipkamo v iskalno okno niz ali zaporedje nizov (ki lahko vsebujejo tudi znaka ? in * za mehko ujemanje), in konkordančnik bo te nize iskal v polju, ki je bilo izbrano na iskalni maski. Polja so lahko sledeča:

1. *word* (beseda) je niz, kot se pojavi v besedilu korpusa, vključno z velikim začetnicami in pregibanjem, npr. *»človeka«, »Človeka«, »Janeza«*
2. *nform* (normalizirana oblika besede) je niz, ki je identičen besedi, samo, da je vedno zapisan z malimi črkami, npr. *»človeka«, »janeza«*
3. *lemma* (lema) je osnovna oblika besede, npr. *»človek«, »janez«*
4. *msd-en* (angleška oblikoskladenjska oznaka) je oznaka SPOOK, podana v angleškem jeziku, npr. »*Ncm*« (za »*Noun common masculine*«)
5. *msd-sl* (slovenska oblikoskladenjska oznaka) je oznaka SPOOK, podana v slovenskem jeziku, npr. »*Som*« (za »*samostalnik občno_ime moški spol*«)
6. *ctag* (korpusna oznaka) je oznaka programa TreeTagger, npr. *»VVINF«* za nemški nedoločnik, ki v sistemu SPOOK ustreza slovenskemu *»Ggn«* oz. *»glagol glavni nedoločnik«* in angleškemu *»Vmn«* oz. »*Verb main infinitive*«.

Z enostavnim načinom tako lahko npr. iščemo vse besede, ki se končajo na *-uje* (iskalni izraz *»*uje«*, polje *»word«*), leme, ki vsebujejo *»gn«* (iskalni izraz *»*gn*«*, polje *»lemma«*), ali pa vsa zaporedja pridevnika in samostalnika (iskalni izraz *»P* S*«*, polje *»msd-sl«*).



Slika 6: Vmesnik konkordančnika CUWI.

Uporaba skladnje CWB v iskalnem oknu je sicer bolj dolgovezna, vendar omogoča izražanje bistveno bolj kompleksnih iskalnih pogojev. Po eni strani lahko tu namesto mehkega ujemanja uporabljamo regularne izraze (npr. *"[^aeiou]+"* za vse pojavnice, ki ne vsebujejo samoglasnikov), po drugi pa lahko medsebojno kombiniramo več polj (npr. *"[msd-sl=«P.*"] [lemma="konj" & ctag ="Somei"]«* za nize, ki so sestavljeni iz pridevnika in besede z lemo »*konj*«, ki je v imenovalniku ednine.

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. v posameznem besedilu. Te omejitve je mogoče vnesti s pomočjo maske programa CUWI ali pa neposredno z ukazi CWB, pri čemer CUWI vedno pokaže, kakšen ukaz je dejansko izvedel, kar uporabniku olajša sestavljanje kompleksnejših iskalnih zahtev.

CUWI podpira več načinov izpisa najdenih fraz v korpusu. Zadetki s kontekstom so lahko prikazani v vezanem besedilu, v standardnem formatu KWIC (»*Key-Word In Context*« oz. ključna beseda s kontekstom) ali pa brez konteksta, vendar s številom pojavitev v korpusu, torej kot frekvenčni seznam. Pri vseh načinih izpisa velja, da lahko izpišemo ne samo pojavnice (besedilo), temveč katerokoli kombinacijo oznak. Konkordančnik tudi ponuja razne možnosti sortiranja (po kontekstu in od zadaj) in filtriranja (vse, naključni vzorec).

Rezultate iskanja lahko shranimo na svoj računalnik kot razpredelnico v formatu CSV ali Excel, kar omogoča nadaljnje analize najdenih nizov v enem od programov za delo z razpredelnicami. Konkordančnik prav tako omogoča shranjevanje rezultatov v obliki podatkovnih struktur za skriptne jezike v formatu JSON (*JavaScript Object Notation*) ali Perl, če jih želimo programsko obdelovati. Kot je bilo že omenjeno, podpira konkordančnik prikaz poravnanih segmentov, če imamo izbran prevedeni slovenski korpus, tudi za več jezikov hkrati. Iskanje je mogoče tudi omejiti s postavitvijo dodatnih iskalnih zahtev nad poravnanimi segmenti. Tako lahko npr. iščemo vse pojavitve leme »*mati*«, kjer se v poravnanih italijanskih segmentih ne pojavi lema »*madre*«.

5 ZAKLJUČKI

Prispevek je predstavil bolj tehnične vidike korpusa SPOOK, in sicer oblikoskladenjsko označevanje in sistem oznak SPOOK, zapis korpusa v XML in TEI in spletni konkordančnik CUWI, s katerim je mogoče iskati po korpusu. Priporočila in vmesnik do konkordančnika so dostopni na spletni strani <http://nl.ijs.si/spook/>, pri čemer so priporočila prosto dostopna po licenci Creative Commons, konkordančnik pa zaradi avtorskih pravic nad izvornimi besedili zahteva za uporabo geslo – zaradi teh omejitev je tudi korpus kot celota dostopen samo partnerjem projekta SPOOK.

Zahvala

Avtor se zahvaljuje Jan Joni Javoršku za implementacijo konkordančnika CUWI in za pripombe na besedilo, za slednje tudi Darji Fišer. Delo predstavljeno v tem delu je financiral projekt ARRS J6-2009-0581 »Slovensko prevodoslovje – viri in raziskave«.

Literatura

Arhar, Špela in Gorjanc, Vojko, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2, 95–110.

Burbeck, Steve, 1992: *Applications Programming in Smalltalk-80(TM): How to use Model-View-Controller (MVC).* <http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html> (Dostop 15. 5. 2012)

Christ, Oliver, 1994: A modular and flexible architecture for an integrated corpus query system. *Proceedings of the Conference in Computational Lexicography, COMPLEX'94*, Hungarian Academy of Sciences, Budimpešta, 23–32.

Creative Commons licence Attribution-ShareAlike 3.0 <http://creativecommons.org/licenses/by-sa/3.0/ > (Dostop 1. 5. 2012)

CWB: The IMS Open Corpus Workbench. < http://cwb.sourceforge.net/> (Dostop 1. 4. 2012)

Erjavec, Tomaž, 2009: Odprtost jezikovnih virov za slovenščino. Stabej, Marko (ur.): *Infrastruktura slovenščine in slovenistike*. 28. Simpozij Obdobja, Ljubljana: Znanstvena založba Filozofske fakultete, 115–121.

Erjavec, Tomaž (ur.), 2010a: *MULTEXT-East Resources Version 4 »MondiLex«*. <http://nl.ijs.si/ME/V4/> (Dostop 1. 5. 2012)

Erjavec, Tomaž, 2010b: Text Encoding Initiative Guidelines and their Localisation. *Infoteka*. 11/1, 3a–14a.

Erjavec, Tomaž, 2012a: MULTEXT-East : morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, 46/1, 131–142.

Erjavec, Tomaž, 2012b: The goo300k corpus of historical Slovene. *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12,* Pariz, ELRA.

Erjavec, Tomaž, Ignat, Camelia, Pouliquen, Bruno in Steinberger, Ralf, 2005: Massive multi-lingual corpus compilation: Acquis Communautaire and ToTaLe. *Proceedings of the 2nd Language & Technology Conference*, Poznan, Poljska, 32–36.

Erjavec, Tomaž, Krek, Simon, 2008: Oblikoskladenjska priporočila in označeni korpusi JOS. Erjavec, Tomaž in Žganec Gros, Jerneja (ur.). Zbornik Šeste konference Jezikovne tehnologije, zbornik 11. mednarodne multikonference Informacijska družba – IS 2008, zvezek C. Ljubljana: Institut Jožef Stefan, 49–53.

Extensible Markup Language (XML), <http://www.w3.org/XML/> (Dostop 1. 4. 2012)

Greer, Derek, 2007: *Interactive Application Architecture Patterns*. <http://aspiringcraftsman.com/2007/08/25/interactive-application-architecture/> (Dostop 15. 5. 2012)

Ide, Nancy in Jean Veronis, 1994: Multext (multilingual tools and corpora). *Proceedings of the 15th International Conference on Computational Linguistics, CoLing'94*, Kyoto, 90–96.

JavaScript Object Notation. <http://www.json.org/> (Dostop 1. 5. 2012)

Jezikovni viri starejšega slovenskega jezika IMP, <http://nl.ijs.si/imp/> (Dostop 1. 5. 2012)

McDaniel, Adam, 2011: Html5: Your Visual Blueprint for Designing Rich Web Pages and Applications. *Visual Blueprint, 7. zv.*, John Wiley & Sons.

Mojolicious – Perl real-time web framework. <http://mojolicio.us/> (Dostop 1. 5. 2012)

Oblikoskladenjske specifikacije JOS V1. <http://nl.ijs.si/jos/josMSD-sl.html> (Dostop 1. 5. 2012)

Projekt JOS: jezikoslovno označevanje slovenskega jezika. <http://nl.ijs.si/jos/> (Dostop 1. 5. 2012)

Schmidt, Douglas C., 1995: Reactor: An Object Behavioral Pattern for Demultiplexing and Dispatching Handles for Synchronous Events. Coplien, Jim in Schmidt, Douglas C. (ur.) *Pattern Languages of Program Design*. Addison-Wesley. Novejša različica: http://www.dre.vanderbilt.edu/~schmidt/PDF/reabctor-siemens.pdf (Dostop 15. 5. 2012)

Schmid, Helmut, 1994: Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.* Manchester, Velika Britanija.

Schmid, Helmut, 1995: Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Irska.

TEI P5: Guidelines for Electronic Text Encoding and Interchange. <http://www.tei-c.org/Guidelines/P5/> (Dostop 1. 4. 2012)

The Unicode Consortium, <http://unicode.org/> (Dostop 1. 4. 2012)

TreeTagger – a language independent part-of-speech tagger. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/> (Dostop 1. 9. 2011)

XSL Transformations (XSLT) Version 1.0, <http://www.w3.org/TR/xslt> (Dostop 1. 4. 2012)