Oblikoskladenjske specifikacije in označeni korpusi JOS

Tomaž Erjavec, Simon Krek

Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si, simon.krek@ijs.si

Povzetek

Jezikovne vire JOS trenutno sestavljajo oblikoslovne specifikacije in dva korpusa. Prvi korpus je "jos100k", enojezični vzorčeni in uravnoteženi korpus slovenskega jezika s 100.000 besedami in z ročno označenimi oz. pregledanimi lemami ter oblikoskladenjskimi oznakami. Drugi je "jos1M", enomilijonski delno ročno pregledani korpus. Oba korpusa sta bila vzorčena iz 620-milijonskega korpusa FidaPLUS. Jezikovni viri JOS so označeni v skladu s označevalnimi standardi, oblikoskladenjske specifikacije skladno s sistemom MULTEXT-East, tako specifikacije kot korpusa pa skladno z navodili združenja Text Encoding Initative (Guidelines P5). Vsi viri so na voljo kot zbirka podatkov za raziskovalne namene po licenci Creative Commons in so namenjeni razvoju jezikovnih tehnologij za slovenski jezik.

The JOS Language Resources: Morphosyntactic Specifications and Annotated Corpora

The JOS morphosyntactic resources for Slovene consist of the specifications and two corpora: jos100k, a 100,000 word balanced monolingual sampled corpus annotated with hand validated morphosyntactic descriptions (MSDs) and lemmas, and jos1M, the 1 million word partially hand validated corpus. The two corpora have been sampled from the 620 million word Slovene reference corpus FidaPLUS. The JOS resources have a standardised encoding, with the MULTEXT-East-type morphosyntactic specifications and the corpora encoded according to the Text Encoding Initiative Guidelines P5. JOS resources are available as a dataset for research under the Creative Commons licence and are meant to facilitate developments of HLT for Slovene.

1. Uvod

Jezikoslovno označeni korpusi so osnovni vir za jezikovne tehnologije, vendar za mnoge jezike še niso na voljo, predvsem kot zaključene podatkovne zbirke. Eden od pomembnih virov so ročno oblikoskladenjsko označeni korpusi, ki so potrebni za učenje oblikoskladenjskih označevalnikov (part-of-speech taggers), ki so sami del osnovne jezikovnotehnološke infrastrukture za nek jezik.

Jezikovni viri za slovenščino, razviti v okviru projekta MULTEXT-East[1] (1995-1997), so vsebovali doslej edini prosto dostopni ročno označeni korpus – slovenski prevod romana "1984" Georgea Orwella. Nabor oznak ter način označevanja korpusov po sistemu MULTEXT-East je bil kasneje uporabljen pri številnih slovenskih korpusih, med drugim pri referenčnem korpusu FIDA (Erjavec in dr., 1998) in FidaPLUS (Arhar in Gorjanc, 2007). Slabost korpusa "1984" pa je njegova velikost (100.000 besed) in predvsem dejstvo, da vsebuje le en preveden roman, kar pomeni, da predstavlja razmeroma skromen vir za učenje oblikoskladenjskih označevalnikov. Poleg tega je večletna raba slovenskega nabora oznak MULTEXT-East pokazala, da bi bile pri naboru potrebne določene spremembe.

Projekt JOS skuša zapolniti vrzel pri jezikovnih virih za slovenščino z izdelavo standardiziranih prosto dostopnih označenih korpusov, skupaj z revidiranim naborom oblikoskladenjskih specifikacij. V članku poročamo o prvih rezultatih: korpusu "jos100k", ki predstavlja zlati standard za označevanje, in korpusu "jos1M", na katerem trenutno poteka delo. Oba korpusa vsebujeta vzorčene odstavke iz korpusa FidaPLUS in sta označena z razdvoumljenimi in ročno preverjenimi lemami in oblikoskladenjskimi oznakami.

[1] http://nl.ijs.si/ME/

2. Vzorčenje korpusa FidaPLUS

Korpus FIDA[2] (Erjavec et al., 1998) je 100-milijonski referenčni korpus sodobne slovenščine in vsebuje besedila, nastala med leti 1990-2000. Označen je v formatu SGML in v skladu s priporočili združenja Text Encoding Initiative TEI P3 (Sperberg-McQueen in Burnard, 1999). Posameznim pojavnicam je pripisan podatek o lemi ter oblikoskladenjski oznaki po sistemu MULTEXT-East (Erjavec, 2004), vendar v primeru večih možnih lem ali oznak te niso razdvoumljene.

Korpus FidaPLUS[3] (Arhar in Gorjanc, 2007) je bistveno večji (620 milijonov besed) ter vsebuje besedila, nastala med leti 1990-2006. Za razliko od korpusa FIDA so leme in oblikoskladenjske oznake v korpusu FidaPLUS razdvoumljene, celoten postopek procesiranja – pretvorba iz izvornega formata, tokenizacija, oblikoskladenjsko označevanje in razdvoumljanje – je bilo izvedeno z programskimi orodji v lasti jezikovnotehnološkega podjetja Amebis.[4] Korpus je prosto dostopen za raziskovalne namene, kot podatkovna zbirka pa je dostopen le članom konzorcija (Amebis, DZS, Univerza v Ljubljani, Univerza v Mariboru, Institut Jožef Stefan), za dostop je namreč potreben podpis konzorcijske pogodbe. Druga omejitev pri uporabi korpusa FidaPLUS za namen učnega korpusa je dejstvo, da je bil v celoti avtomatično označen. Označevalnik podjetja Amebis označuje s približno 85-odstotno natančnostjo, ostalih 15 % poleg napak pri označevanju vključuje tudi neoznačene neprepoznane besede, ki jih je približno 2 %. Kljub temu korpus predstavlja dobro podlago za razvoj prosto dostopnega ročno označenega korpusa za namene jezikovnotehnoloških raziskav.

[2] http://www.fida.net/
[3] http://www.fidaplus.net/
[4] http://www.amebis.si/

Prvi korak na poti od korpusa FidaPLUS do korpusa JOS je bila pretvorba v format XML po priporočilih TEI P4 (Sperberg-McQueen in Burnard, 2002), da bi s tem ohranili standardni format in omogočili uporabo orodij za delo s formatom XML, predvsem XSLT. Format TEI P4 je sicer povratno združljiv s formatom TEI P3 korpusa FIDA in XML je podmnožica formata SGML, vendar končni korpus FidaPLUS kot podatkovna zbirka ni v celoti skladen niti s formatom SGML niti s specifikacijami MULTEXT-East, zato je bil proces pretvorbe zahtevnejši, kot je bilo pričakovati. Procesiranje je bilo izvedeno s pomočjo niza skript v programskem jeziku Perl, končni pretvorjeni korpus FidaPLUS pa imenujemo Fida+X. Ta je za malenkost manjši kot izvorni korpus, ker smo izpustili besedila, pri katerih hevristični postopki s pomočjo Perl skript niso zadostovali za njihovo kompatibilnost s standardom TEI P4. Korpus Fida+X je bil uporabljen kot vir za izdelavo korpusov JOS.

2.1. Postopek vzorčenja

Korpusa jos100k in jos1M sta nastala iz korpusa Fida+X z dvostopenjskim filtriranjem in procesom vzorčenja z namenom, da pri končnem rezultatu dosežemo naslednje cilje:

- Uravnoteženost in reprezentativnost: slednja lastnost izhaja iz zasnove korpusa FidaPLUS, pri čemer pa ta vsebuje velik delež časopisnih besedil in relativno majhen delež literarnih, predvsem pa strokovnih besedil. Ob preprostem prevzemanju razmerij iz korpusa FidaPLUS bi bila uravnoteženost korpus JOS vprašljiva.
- Kvaliteta besedil: ker bo končni korpus jezikoslovno označen, je bilo pomembno, da vsebuje le smiselne in zaključene odstavke in pojavnice: FidaPLUS vsebuje dele besedil, ki se ponavljajo, in primere, kjer so ob pretvorbi iz izvornih besedil nastali zelo kratki stavki ali odstavki, ki včasih vsebujejo tudi ostanke podatkov o formatiranju.
- Avtorske pravice: čeprav bi bilo za določene tipe analiz bolje vključiti celotna besedila, bi bilo to vprašljivo s stališča kršenja avtorskih pravic besedilodajalcev, vključitev krajših delov besedil pa je manj problematična – postopek vzorčenja poleg tega prispeva k večji raznolikosti korpusa.

Pri prvem koraku vzorčenja so naključna izbrana celotna besedila iz korpusa Fida+X, vendar so izključena besedila z nepravilnimi deli, določeni tipi besedil pa so v vzorcu bolj zaželeni. Filter najprej izključi besedila, ki so prekratka ali predolga, vsebujejo podatke o formatiranju ali so slabo oblikovana glede na različne hevristične kriterije. Sledi faza obteževanja po tipih besedil in drugih metapodatkih z namenom, da se uravnoteži delež časopisnih besedil v primerjavi z drugimi tipi besedil. V drugem koraku postopka so izbrani naključni odstavki, ki so ponovno podvrženi preverjanju: minimalni ali maksimalni dolžini ter enkratnosti pojavljanja. Oba koraka sta bila izvedena dvakrat, vsakič z drugimi nastavitvami. Za korpus jos100k je bil za prvi korak izdelan 10-milijonski korpus, za jos1M pa 100-milijonski. Za oba korpusa je bil potem v drugem koraku izbran en odstotek odstavkov.

Tabela 1 vsebuje podatke o številu besedil, odstavkov, stavkov, besed ter pojavnic (besed skupaj z ločili).

Korpus	jos100k	jos1M
<text>	249	2.565
<p>	1.599	15.758
<s>	6.151	60.291
<w>	100.003	1.000.019
<w> + <c>	118.394	1.182.945

Tabela 1: Število oznak v obeh korpusih JOS.

V korpusu FidaPLUS je vsako besedilo označeno glede na tip in zvrst besedila. Pri prvem koraku vzorčenja korpusov JOS smo te kategorije različno obtežili in razmerja se tako razlikujejo od korpusa FidaPLUS. V tabeli 2 in 3 navajamo razmerja med vrhnjimi kategorijami v obeh korpusih JOS. Razmerja smo med prvim in drugim vzorčenjem rahlo spremenili, zato se pri obeh razlikujejo. JOS1M vsebuje manj literarnih besedil in časopisnega gradiva in več naravoslovnih in tehničnih besedil ter mesečnih publikacij.

Zvrst	jos100k	jos1M
umetnostna besedila (proza)	10,1 %	6,7 %
neumetnostna besedila (nestrokovna)	67,6 %	66,6 %
družboslovje in humanistika	9,6 %	9,9 %
naravoslovje in tehnologija	6,5 %	13,3 %
skupaj	93,8 %	96,6 %

Tabela 2: Razmerja med besedili po zvrsteh

Tip	jos100k	jos1M
monografija	28,1 %	22,9 %
mesečna revija	9,6 %	16,2 %
štirinajstdnevna revija	2,3 %	2,0 %
tedenska revija	7,3 %	9,1 %
tedenski časopis	9,0 %	10,4 %
dnevni časopis	37,7 %	24,3 %
skupaj	94,1 %	95,0 %

Tabela 3: Razmerja med besedili po tipu

3. Oblikoskladenjske specifikacije in nabor oznak JOS

Oblikoskladenjske specifikacije MULTEXT(-East) temeljijo na delu skupine EAGLES (Calzolari in Monachini, 1996) in določajo strukturo in vsebino veljavnih oblikoskladenjskih oznak ali MSD-jev (morpho-syntactic descriptions). Specifikacije za vsak posamezen jezik opredeljujejo, katere so veljavne oznake in kaj pomenijo. Tako na primer določajo, da je MSD s črkovnim nizom `Sometn` veljaven za označevanje slovenščine in je ekvivalenten naboru naslednjih lastnosti: `samostalnik, vrsta = občni, spol = moški, število = ednina, sklon = tožilnik, živost = ne`. Ker je slovenščina oblikoslovno izjemno bogat jezik z velikim številom lastnosti pri pregibnih besednih

vrstah, je število veljavnih oznak precej večje kot pri večini zahodnoevropskih jezikov – okrog 2.000.

Izhodišče pri odločanju o prenovi nabora oznak MULTEXT-East je bila ocena, da osnovni način formalnega zapisa in struktura oznak dobro služi svojemu namenu, da pa prihaja do težav pri določenih lastnostih in njihovih vrednostih, predvsem pri nekaterih dovoljenih kombinacijah lastnost-vrednost, ter pri pripisovanju nekaterih MSD-jev določenim lemam in oblikam. Nadaljnja težava, tokrat celotnega nabora specifikacij MULTEXT-East, je razvrstitev lastnosti v črkovni niz MSD-jev. Ker specifikacije veljajo za celo vrsto različnih jezikov, tiste lastnosti, ki so značilne samo za določen jezik, končajo na koncu črkovnega niza s praznimi mesti pri lastnostih, ki jih jezik ne izkazuje. Tako lahko pride do izjemno dolgih črkovnih nizov, kot je npr. glagolski `Gppspe--n-----d`. Smiselno je torej dovoliti prerazporeditev mesta lastnosti glede na posamezen jezik, kar omogoči, da so nizi krajši, hkrati pa s pomočjo preslikave lastnosti in vrednosti ohranimo kompatibilnost označevanja z drugimi jeziki.

Namen prenove oblikoskladenjskih specifikacij je bil med drugim tudi standardizacija nabora oznak za slovenščino. Zato je bila opravljena analiza označevanja korpusa FidaPLUS ter razmeroma obsežen pregled drugih naborov oznak za slovenščino ter za druge jezike. Pri slovenskem jeziku sta bila upoštevana nabora oznak, uporabljena pri označevanju korpusa LC-STAR (Verdonik et al., 2004) ter korpusa Nova beseda (Jakopin in Bizjak, 1997). Zadnji se od prvih dveh temeljno razlikuje, saj ne uporablja pozicijskega načela pri pripisovanju lastnosti in se močno opira na tradicionalni slovnični opis jezika (Lönneker, 2005). Od drugih jezikov so bili podrobneje analizirani nabori CLAWS za angleški jezik ter češki nabor AJKA ter nabor oznak, uporabljen pri označevanju Češkega nacionalnega korpusa[5] ter Praške odvisnostne drevesnice.[6]

Končni nabor oznak JOS v osnovi ohranja načela nabora MULTEXT-East, vendar postopek preslikave med obstoječimi oznakami v korpusu FidaPLUS po sistemu MULTEXT-East in novimi po sistemu JOS ni trivialen, kajti pri pretvorbi je potrebno upoštevati tudi besedno obliko in/ali lemo.

Tabela 4 kaže povzetek kategorij in lastnosti nabora oznak JOS. Specifikacije vsebujejo tudi listo veljavnih MSD-jev, ki jih je 1.908, pri vsakem pa so dodani tudi konkretni primeri besednih oblik iz leksikona. Nabor oznak je torej kljub spremembam precej obsežen in nekatere kombinacije lastnosti so še vedno precej redke. Tako denimo se v korpusu jos100k pojavlja 1.064 različnih MSD-jev, kar je le 55,7 % celotnega nabora.

Slika 1 kaže primer iz specifikacij nabora JOS v formatu XML. Specifikacije so na voljo v angleškem in slovenskem jeziku, to velja za kategorije, lastnosti in vrednosti. S pomočjo specifikacij in ustrezne datoteke XSLT je tako denimo mogoče prevesti slovensko oznako `Sometn` v angleško `Ncmsan` (`Noun, Type = common, Gender = masculine, Number = singular, Case = accusative, Animate = no`), kar omogoča enostavno prehajanje med oznakami v angleškem in slovenskem jeziku.

besedna vrsta	lastnosti s številom vrednosti
samostalnik	vrsta(2), spol(3), število(3), sklon(6), živost (2)
glagol	vrsta(2), vid(3), oblika(7), oseba(3), število(3), spol(3), nikalnost(2)
pridevnik	vrsta(3), stopnja(3), spol(3), število(3), sklon(6), določnost(2)
prislov	stopnja(3), deležje(2)
zaimek	vrsta(9), oseba(3), spol(3), število(3), sklon(6), število_svojine(3), spol_svojine(3), oblika(2)
števnik	zapis(3), vrsta(4), spol(3), število(3), sklon(6), določnost(2)
predlog	sklon(6)
veznik	vrsta(2)
členek	brez lastnosti
medmet	brez lastnosti
okrajšava	brez lastnosti
neuvrščeno	vrsta(3)

Tabela 4: Kategorije nabora oznak JOS z lastnostmi in številom vrednosti

```
<div type="section" xml:id="msd.N">
<head>SAMOSTALNIK</head>
<table n="msd.cat" xml:id="msd.cat.N">
<head xml:lang="sl">Tabela atributov in
vrednosti za samostalnik</head>
<head xml:lang="en">Attribute-value table for
Noun</head>
<row role="type">
<cell role="position">0</cell>
<cell role="name"
xml:lang="sl">samostalnik</cell>
<cell role="code" xml:lang="sl">S</cell>
<cell role="name" xml:lang="en">Noun</cell>
<cell role="code" xml:lang="en">N</cell>
</row>
<row role="attribute">
<cell role="position">1</cell>
<cell role="name" xml:lang="sl">vrsta</cell>
<cell role="name" xml:lang="en">Type</cell>
<cell role="values">
<table>
<row role="value">
<cell role="name" xml:lang="sl">občno
ime</cell>
<cell role="code" xml:lang="sl">o</cell>
<cell role="name" xml:lang="en">common</cell>
<cell role="code" xml:lang="en">c</cell>
</row>
<row role="value">
<cell role="name" xml:lang="sl">lastno
ime</cell>
<cell role="code" xml:lang="sl">l</cell>
<cell role="name" xml:lang="en">proper</cell>
<cell role="code" xml:lang="en">p</cell>
</row>
</table>
</cell>
</row>
...
```

Slika 1: oblikoskladenjske specifikacije JOS – začetek tabele za samostalnik

[5] http://ucnk.ff.cuni.cz/

[6] http://ufal.mff.cuni.cz/pdt/

4. Jezikoslovno označevanje korpusa

Ročno označevanje korpusov je izvedla ekipa študentov pod strokovnim nadzorom. Študenti so preverjali in popravljali oznake, ki so bile iz izvornih oznak po naboru MULTEXT-East iz korpusa FidaPLUS preslikane v nabor JOS. Ročno označevanje je potekalo s pomočjo spletnega vmesnika, ki na podlagi danih parametrov (npr. regularnih izrazov, ki lahko upoštevajo besedno obliko, lemo ali MSD) generira preglednice v formatu MS Excel. Preglednica vsebuje list z osnovnimi podatki o vsebini, list z besedilom in oznakami, ki jih je potrebno pregledati, ter list z navodili. Po ročnem preverjanju preglednico na isti spletni strani naložimo nazaj v korpus, pri čemer je korpus avtomatsko obnovljen z novimi ročno pregledanimi oznakami. Sistem je bil izvorno razvit za popravljanje in označevanje zgodovinskih besedil (Erjavec, 2007) in je bil uspešno uporabljen tudi v projektu JOS. Proces označevanja je cikličen, z mešanimi ročnimi in avtomatskimi postopki.

4.1. Označevanje korpusa jos100k

Označevanje korpusa jos100k je potekalo vzporedno z nastajanjem oblikoskladenjskih specifikacij JOS, novega nabora oznak in preslikave MSD-jev. Proces je bil zato bolj zapleten, vendar so specifikacije, nabor oznak in oznake v korpusu zato konsistentnejši. Celoten korpus jos100k sta preverjala po dva različna označevalca, pojavnice, pri katerih je med njima prihajalo do razlik, pa so bile preverjene še s strani tretjega. V primerih, kjer so bile odkrite nekonsistentnosti pri označevanju določenih kategorij oz. lastnosti, so bile te v kasnejših korakih odpravljene, korpus pa poenoten, rezultat pa je bil v celoti ročno označeni zlati standard JOS.

4.2. Označevanje korpusa jos1M

Ker finančna shema projekta ne dopušča ročnega preverjanja celotnega milijonskega korpusa JOS, se pri milijonskem korpusu jos1M poleg prvega koraka avtomatske pretvorbe iz nabora oznak MULTEXT-East v nabor JOS ročno preverjajo le "sumljive" oznake, v nadaljevanju pa opisujemo uporabljeno metodologijo.

Na ročno označenem korpusu jos100k smo izvedli učni proces z označevalnikom TnT (Brants, 2000) in mu pri postopku kot oporo dodali leksikon besednih oblik, izdelan iz celotnega korpusa Fida+X, skupaj s konvertiranim naborom oznak JOS. V korpusu jos1M ročno preverjamo le pojavnice, kjer se oznake, ki jih je pripisal označevalnik TnT, ter že obstoječe (pretvorjene) oznake iz korpusa Fida+X, ki jih je pripisal Amebisov označevalnik, razlikujejo.

Da bi preverili točnost obeh označevalnikov in prekrivanje napak, ki jih delata, smo izvedli poskus z desetkratnim navskrižnim preverjanjem na korpusu jos100k z označevalnikom TnT, kar pomeni, da smo učili označevalnik na 90 % besedila in označili preostalih 10 %, postopek pa smo ponovili desetkrat na vseh desetinah korpusa. Rezultat je bil korpus, kjer je vsaka pojavnica označena s tremi oznakami: ročno preverjena oznaka, oznaka, ki jo je pripisal Amebisov označevalnik in oznaka, ki jo je pripisal označevalnik TnT. V tabeli 5 je prikazan rezultat poskusa.

V prvi vrstici podamo celoten obseg korpusa glede na število besed. Druga vrstica kaže oceno natančnosti označevalnika TnT (86,6%), tretja pa oceno natančnosti Amebisovega označevalnika (85,7%). Amebisov označevalnik sicer ne pripisuje oznak neznanim besedam, ki obsegajo približno dva odstotka pojavnic, pri ročnem označevanju pa so bile označene vse pojavnice (tudi npr. tujejezične citatno pisane pojavnice, ki spadajo v kategorijo "neuvrščeno"), enako tudi z označevalnikom TnT. Četrta vrstica zajema primere, ko sta oba označevalnika pripisala pravilni MSD (78%), v naslednjih štirih vrsticah pa podajamo rezultat za primere, ko sta se bodisi en ali drugi ali oba označevalnika zmotila. Predvsem je pomemben rezultat v zadnji vrstici, kjer sta oba označevalnika pripisala enako oznako, vendar oba napačno. Pri ročnem preverjanju korpusa jos1M namreč zajemamo le tiste pojavnice, kjer se označevalnika pri pripisu oznake ne strinjata. Pri milijonskem korpusu izračun 7,7% + 8,6% + 2,4% pokaže, da je potrebno ročno validirati okoli 190.000 pojavnic, končna pravilnost označevanja pa je ocenjena na 96,8 %, saj moramo odšteti primere (3,2%), kjer s križanjem rezultatov obeh označevalnikov ne moremo zaznati napak.

Ročno označevanje milijonskega korpusa je v teku, hkrati pa raziskujemo načine, kako izboljšati rezultat pri označevanju izvornih besedil iz korpusa Fida+X s kombiniranjem rezultatov označevanja obeh označevalnikov.

	Št. besed	Ročno	Amebis	TnT	Razlaga
1	100,003	MSD1			Besed v korpusu jos100k
2	86,617	MSD1		MSD1	TnT pravilno označenih
3	85,719	MSD1	MSD1		Amebis pravilno označenih
4	78,011	MSD1	MSD1	MSD1	Oba pravilno označila
5	7,708	MSD1	MSD1	MSD2	Amebis pravilno, TnT narobe
6	8,606	MSD1	MSD2	MSD1	Amebis narobe, Tnt pravilno
7	3,238	MSD1	MSD2	MSD2	Oba narobe, in enako
8	2,440	MSD1	MSD2	MSD3	Oba narobe, in različno

Tabela 5: natančnost označevanja korpusa jos100k – označevalnik Amebis / TnT

5. Format korpusov

Tako korpusa kot oblikoslovne specifikacije so kodirani v formatu XML v skladu s priporočili združenja Text Encoding Initative. Čeprav je bila pri razvoju korpusov zaradi kompatibilnosti s korpusom FidaPLUS uporabljena inačica priporočil P4, je za javno dostopno inačico uporabljena zadnja izdaja priporočil, TEI P5 (TEI Consortium, 2007). Rezultati so dostopni skupaj s pripadajočo shemo XML, ki omogoča formalno validiranje specifikacij in korpusov.

Korpusa sta sestavljena iz dveh delov. Kolofon (TEI header) vsebuje metapodatke, kjer se med drugim nahajajo podatki o velikosti korpusa, o uporabi oznak, naštete so odgovorne osebe, bibliografski podatki o vsebovanih besedilih, vsebuje pa tudi celoten nabor

oblikoslovnih oznak nabora JOS in njihovo dekompozicijo na pare lastnost-vrednost. Drugi del korpusa sestavljajo besedila.

Vsako besedilo v korpusu vsebuje povezavo na opis v kolofonu in je sestavljeno iz niza vzorčenih odstavkov, te sestavljajo stavki in stavke posamezne pojavnice in ločila. Poseben element označuje presledke, vsaka pojavnica pa ima v obliki atributa podatek o lemi in oblikoslovni oznaki.

6. Dostopnost

Jezikovni viri JOS so dostopni na spletni strani projekta,[7] korpuse pa je dovoljeno prenesti na svoj računalnik in uporabljati v skladu z licenco Creative Commons: Priznanje avtorstva-Nekomercialno.[8] Komercialna uporaba korpusov ni dovoljena, ker tega ne dopušča pogodba med besedilodajalci in konzorcijem partnerjev, ki so izdelali korpusa FIDA in FidaPLUS.

Korpusa JOS sta na voljo v izvornem formatu XML TEI P5 ter v več formatih, ki so primerni za predprocesiranje in druge predelave. Med njimi je omembe vreden predvsem format za Corpus Workbench[9] (Christ, 1994), pri katerem vsaka vrstica vsebuje bodisi formalno strukturno oznako ali s tabulatorjem ločeno informacijo o besedni obliki, lemi in MSD-ju, skupaj z razčlenjenimi lastnostmi. To omogoča iskanje po posameznih lastnostih pojavnice ne glede na najvišjo kategorijo, besedno vrsto. Tako bi pri iskanju vseh pojavnic, ki označene z vrednostjo "ženski spol" in "rodilnik", v iskalniku Corpus Workbench uporabili iskalni pogoj [spol="ženski" & sklon="rodilnik"]. Korpusi so dostopni tudi preko spletnega vmesnika, ki uporablja Corpus Workbench, in je ravno tako dosegljiv na domači strani projekta.

7. Zaključek

V prispevku smo predstavili prve rezultate projekta JOS, predvsem dokončani korpus jos100k in korpus jos1M, ki je v zaključni fazi izdelave. Prispevek je opisal korpus FidaPLUS kot vir za oba korpusa JOS, postopke čiščenja in vzorčenja besedil, prenovljen nabor oznak JOS ter oblikoslovnih specifikacij. Komentiral je tudi postopek jezikoslovnega označevanja korpusov, njun format in dostopnost. Predstavljena korpusa sta prva kvalitetno označena in brezplačno dostopna jezikovna vira za slovenščino in bosta znatno olajšala raziskave pri avtomatskem oblikoskladenjskem označevanju in lematizaciji slovenskih besedil.

Nadaljnje delo pri projektu JOS zajema naslednji dve ravni jezikoslovnega označevanja besedil, predvsem skladenjsko označevanje korpusa jos100k in pomensko označevanje in razdvoumljanje s pomočjo slovenskega semantičnega leksikona, narejenega po vzoru WordNet (Erjavec in Fišer, 2006).

Zahvala

Avtorja se zahvaljujeta recenzentom za koristne pripombe. Delo opisano v tem prispevku sta omogočila projekt ARRS J2-9180 “Jezikoslovno označevanje slovenskega jezika: metode in viri” in projekt EU 6FP-033917 SMART “Statistical Multilingual Analysis for Retrieval and Translation”.

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

[8] http://creativecommons.org/licenses/by-nc/2.5/si

[9] http://cwb.sf.net

Literatura

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

Brants T. (2000). TnT - A Statistical Part-of-Speech Tagger. V Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000 (str. 224–231). ACL.

Calzolari, N. in Monachini, M. (ur.) (1996). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: A common proposal and applications to European languages. EAGLES Report EAG—CLWG—MORPHSYN/R. Pisa: ILC.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. V Proceedings of COMPLEX ’94 (str. 23–32). Budimpešta.

Erjavec, T. in Fišer, D. (2006). Building Slovene WordNet. V Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006 (str. 1678–1683). Pariz: ELRA.

Erjavec, T., Gorjanc, V. in Stabej, M. (1998). Korpus FIDA. V Proceedings of the Conference ’Language Technologies for the Slovene Language’ (str. 124–127). Ljubljana: IJS.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004 (str. 1535–1538). Pariz: ELRA.

Erjavec, T. (2007). An architecture for editing complex digital documents. V Proceedings of INFuture2007: “Digital Information and Heritage” (str. 105-114) Zagreb: Fakulteta za humanistiko in socialne vede**.**

Jakopin, P. in Bizjak, A. (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična revija 45 (3–4), 513–532.

Lönneker, B. (2005). Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo? Slavistična revija 53 (2), 193–210.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman A. in Divjak, D. (2008). Designing and evaluating a Russian tagset. V Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2008. Pariz: ELRA.

Sperberg-McQueen, C. M. in Burnard, L. (ur.) (1999). Guidelines for Electronic Text Encoding and Interchange Revised Reprint. The TEI Consortium.

Sperberg-McQueen, C. M. in Burnard, L. (ur.) (2002). Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines. The TEI Consortium.

TEI Consortium (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Verdonik, D., Rojc, M. in Kačič, Z. (2004). Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004. Pariz: ELRA.