Tomaž Erjavec
Skupina za jezik in govor, E8, Institut Jožef Stefan

UDK 81:681.3

Računalniške zbirke besedil

1 Uvod

Korpus je zbirka besedil, ki so izbrana tako, da karakterizirajo stanje ali raznovrstnost nekega jezika. Uporaben je kot osnova, na kateri gradimo opise jezika, ali pa kot sredstvo za preverjanje hipotez o jeziku. Korpusi so dandanes že standardno shranjeni na računalnikih, saj ti po eni strani omogočajo kompaktno in poceni hranjenje ter razpečevanje ogromnih količin besedil, po drugi strani pa ta besedila lahko z njimi bolj učinkovito izkoriščamo. Uporabnost nekega korpusa je odvisna od njegove velikosti pa tudi urejenosti, tj., kako podrobno je dokumentiran in označen, ter standardiziranosti njegovega zapisa.

Veja jezikoslovja, ki je korpuse tradicionalno uporabljala, je leksikografija; pri izdelavi slovarjev metode introspekcije ne zadoščajo in se je nujno opreti na govor (parole). V formalnem in računalniškem jezikoslovju ta pristop ni nujno edini.

1.1 Nekaj zgodovine

Računalniški korpusi in (predvsem kvantitativne) obravnave le-teh so bile popularne že v petdesetih in šestdesetih letih, nato pa so doživele zaton, predvsem zaradi velikega vpliva teorij N. Chomskega. Pretvorbeno-tvorbena slovnica in njene naslednice namreč jemljejo kot predmet preučevanja »notranji jezik«, tj. človeško sposobnost produkcije jezika. Kriterij za ustreznost primerov, ki jih ti pristopi obravnavajo, je občutek govorcev o njihovi pravilnosti. Zbirke jezika so tako manj zanimive, saj vsebujejo napake in moteče elemente, po drugi strani pa relativno malo teoretično zanimivih primerov. Ne samo v formalnem, temveč tudi v računalniškem jezikoslovju je od konca šestdesetih in približno do sredine osemdesetih let v ospredju zanimanje za formalizacije jezika, ki temeljijo na pravilih in udejanjajo idealizirano znanje govorcev jezika.

Dejavnikov, ki so v osemdesetih letih vplivali na ponovni prodor empirično podprtega jezikoslovja, je več. Programi za skladenjsko analizo so sicer lahko minuciozno razčlenili neki točno določen stavek, vendar pa so dosegali zelo slabe rezultate na odprtem besedilu. Razlog za to je bil predvsem v premajhnem pokritju njihovih slovarjev in pravil, pri čemer pa je izdelava teh podatkov izredno zamudno pa tudi zahtevno delo. Ta problem, t. i. »knowledge acquisition bottleneck«, je tipičen ne samo za računalniško jezikoslovje, pač pa tudi za večino področij umetne inteligence. Postalo je jasno, da je za napredek tega področja potrebno začeti zajemati vire informacij o domeni obravnave (npr. o onesnaženosti jezer ali diagnozah pacientov) in se na njihovi osnovi (pol)avtomatsko učiti zakonitosti, ki v tej domeni vladajo. V računalniškem jezikoslovju so takšni *jezikovni viri* še posebej kompleksni in je njihovo zbiranje temu primerno težje, s čimer postane tudi rezultat toliko pomembnejši. Zbiranje se je osredotočilo na bolj ali manj formalno zapisane računalniško berljive slovarje, predvsem pa na korpuse kot osnovne vire jezika.

Obenem je nova generacija programov, ki temeljijo na statističnih zakonitostih besedila, pokazala obetavne rezultate. Ti programi so po svoji naravi sicer nepopolni, so pa bolj robustni in imajo v povprečju precej večje pokritje od simboličnih pristopov, poleg tega pa se lahko učijo iz primerov. Lažje je ročno označiti neko besedilo, na katerem se bo program učil, kot pa pisati pravila, ki naj bi te označbe zagotovila. Poleg tega je ročno označen ali pregledan jezikovni vir lahko koristen tudi

za druge namene, ročno napisana pravila pa uporabna samo skupaj s programom, za katerega so bila napisana.

To nas pripelje do pomembne razlike med računalniškimi korpusi petdesetih let in sedanjimi korpusi. V pedesetih letih so bila besedila tipično zbrana za neki povsem določen namen in v formatu, ki ga je podpirala programska oprema, ki naj bi besedilo obdelala. Ker jezikovni viri danes pomenijo dragoceno blago, ki ga je potrebno ohraniti pa tudi širiti, se v njihovo izdelavo vlaga več truda, zapisuje pa se jih v skladu z mednarodnimi standardi in priporočili.

Nenazadnje je bliskoviti dvig količine in kvalitete računalniških korpusov pripisati tudi tehnološkemu napredku na področju računalništva in z njim spremembi glavne namembnosti računalnikov. Računalniki se vedno bolj uporabljajo kot orodje za procesiranje besedil, s čimer postaja jezikovni inženiring profitno področje, obenem pa se začenja vprašanje »računalniške pismenosti« nekega jezika povezovati z njegovo identiteto. Programi, ki naj bi pomagali pri pripravi, izmenjevanju, urejanju, predstavitvi in dostopu do jezikovnih informacij za neki jezik, tipično potrebujejo urejene vire znanja o tem jeziku. Do takšnih virov najlaže pridemo s pomočjo besedilnih zbirk. Obenem je vse več besedil dostopnih neposredno na računalnikih in jih je temu primerno lažje pretvoriti v korpus.

V ilustracijo napredka računalniških korpusov lahko primerjamo velikost prvega označenega referenčnega korpusa z današnjimi korpusi (britanske) angleščine. Korpus LOB (Lancaster-Oslo/Bergen) [6], izdelan leta 1986, je vseboval milijon besed, korpus BNC (Brittish National Corpus), izdelan leta 1994, pa sto milijonov besed — v tiskani obliki bi ta besedila zavzela približno deset metrov polic. Na količino korpusov in zanimanje zanje kaže tudi ustanovitev 'borkerskih hiš' za korpuse in druge jezikovne vire. Tako je bil leta 1992 v Združenih državah z vladno podporo ustanovljen Linguistic Data Consortium, ki združuje v svoji ponudbi preko štirideset pisnih in govorjenih korpusov ter slovarskih baz. Pred nedavnim je tej pobudi sledila tudi Evropska unija s financiranjem ustanovitve organizacije ELRA (European Linguistic Resources Association).

Količina in raznovrstnost jezikovnih virov je seveda največja za anlgeški jezik. V zadnjih desetih letih je bilo mnogo takšnih virov, na prvem mestu korpusov, izdelanih tudi za jezike Evropske unije, k čemur so v veliki meri prispevale tudi iniciative Evropske unije. Za jezike vzhodno- in srednjeevropskih držav je stanje slabše in obenem precej raznovrstno. Določene države imajo na področju (računalniškega) jezikoslovja že dolgo tradicijo (npr. Češka in Madžarska), kar se odraža tudi v stanju njihove jezikovne infrastrukture. Tako imajo npr. na filozofski fakulteti v Pragi že oddelek, katerega edina naloga je zagotoviti 20 milijonov besed velik referenčni korpus, ki bo nato podlaga novemu slovarju češkega jezika.

V Sloveniji dostopnih in obenem standardiziranih jezikovnih virov še nimamo. Edini javni referenčni korpus slovenskega jezika [16] je precej majhen, obstaja samo v knjižni obliki in je star dvajset let. Svetla izjema tega stanja so na WWW objavljena besedila slovenskih klasikov [9], ki imajo prednost, da so dostopna in že do precejšnje mere urejena, ravno tako pa, glede na svojo starost, večinoma ne podležejo več zakonu o avtorskih pravicah. Vendar pa ravno zaradi starosti besedil ne podajajo slike sodobnega slovenskega jezika. Dosti slovenskih besedil, ki bi bila lahko osnova za korpuse, obstaja seveda tudi v računalniški obliki, vendar pa niso standardizirana, predvsem pa je njihova dostopnost omejena na institucije, ki so jih proizvedle (npr. založbe, časopisne hiše), ali pa na institucije, ki so sodelovale v njihovi pripravi.

1.2 Tipologija korpusov

S širjenjem računalniških korpusov se je pojavila tudi potreba po ovrednotenju ter razvrstitvi korpusov. Z opisom karakteristik, s katerimi lahko neki korpus ovrednotimo, in z definiranjem zvrsti korpusov, ki jih je med seboj smiselno razlikovati, se je ukvarjala skupina za tipologijo korpusov pri evropski iniciativi Eagles (Expert Advisory Group for Language Engineering).

Po tipologiji Eagles [14] so karakteristike nekega korpusa naslednje:

— *velikost*, tj. količina podatkov, ki jih neki korpus vsebuje;
— *kakovost* njegove izdelave;
— *avtentičnost* glede na kriterije, po katerih je bil zgrajen;
— *enostavnost* njegovega zapisa;
— *dokumentiranost*.

Zvrsti korpusov pa so:

— *referenčni korpusi*, ki predstavljajo osnovno zvrst korpusa in služijo kot jezikovni standardi. Posebna pozornost se pri takšnih korpusih posveča izbiri komponentnih besedil, saj naj bi korpus predstavljal idealizirano podobo nekega jezika. Primer takšnega korpusa je prvi široko dostopni računalniški korpus, in sicer korpus Brown ameriške angleščine [12], ki vsebuje 500 skrbno uravnoteženih odlomkov iz petnajstih jezikovnih zvrsti, ki segajo od religije, preko znanstvene fantastike, do humorja. Referenčne korpuse kontrastiramo s *specializiranimi korpusi*, ki služijo nekemu točno dočenemu namenu, in *oportunističnimi* korpusi, ki so zbrani glede na dane možnosti in služijo kot cenena inačica referenčnih korpusov;
— *govorjeni in govorni korpusi* vsebujejo, za razliko od *pisnih korpusov*, govor oziroma transkripcijo govora; lahko bi trdili, da so edino takšni korpusi avtentični, saj se jezik primarno konstituira skozi govor in ne pisano besedo. Takšni korpusi se v velikih količinah pojavljajo šele v zadnjem času, predvsem zato, ker so zanimivi za avtomatsko procesiranje govora kot enega bolj prodornih področij jezikovnih tehnologij;
— *korpusi podjezikov* so omejeni in specializirani, saj zajemajo jezik v točno določeni funkciji. Vsebujejo npr. tehnične priročnike nekega področja ali pa posnetke dialogov med piloti in stolpom na letališčih. Taki korpusi so ponavadi tudi zbrani za določen namen;
— *vzorčni korpusi* niso sestavljeni iz celotnih besedil, temveč iz fragmentov besedil. Tako je npr. korpus Brown sestavljen iz enakomerno dolgih odlomkov po dva tisoč besed. Razlogi za izdelavo vzorčnega namesto *celostnega korpusa* so predvsem zgodovinske ali pa pravne narave. Kapacitete računalnikov so bile včasih dosti manjše, pa tudi računalniško berljivo besedilo je bilo težje dostopno. Poleg tega pa lastniki besedil prej kot vključitev celotnih besedil dopustijo vključitev fragmentov svojih besedil v neki korpus, saj s tem otežijo možnost neavtoriziranega ponatisa teh besedil. Seveda pa je vzorčni korpus manj kvaliteten od celostnega, saj ne podaja integralne podobe besedil, ki jih zajema;
— *spremljevalni korpusi* so, za razliko od klasičnih korpusov, dinamični. Jezik se spreminja, in korpus, izdelan danes, že jutri ne odraža trenutne podobe jezika. Ker je vedno več besedil dostopnih neposredno v računalniško berljivi obliki, postaja zajem besedil lažji, s tem pa rastejo tudi možnosti za vzdrževanje te, še precej nove zvrsti korpusov;
— *primerljivi korpusi* vsebujejo primerljiva besedila v več jezikih, npr. časopisne članke iz evropskih časopisov v nekem obdobju. Takšni korpusi so koristni za prevajalske študije;
— *vzporedni korpusi* so primerljivi korpusi, ki vsebujejo besedila in njihove prevode. Takšni korpusi so, posebej še za prevajalske študije, jezikovni vir par excelence, predvsem za izdelovanje dvo- in večjezičnih slovarjev. Vendar pa je takšna vzporedna besedila, razen za omejena področja, težko zagotoviti.

1.3 Uporabnost

In kje so korpusi pravzaprav uporabni? Najbolj evidentno področje je seveda slovaropisje. Prvi slovar, izdelan izključno na osnovi računališkega korpusa, je bil Collinsov CoBuild English Language Dictionary iz leta 1987 [13]. Danes je uporaba računalniških korpusov v angleških leksikografskih hišah že standardna, posebej še za specializirane slovarje. Tako npr. v Cambridge

University Press pri izdelavi učnega špansko-angleškega slovarja uporabljajo korpus popravljenih nalog španskih učencev angleščine, saj te najbolje pokažejo tipične napake, na katere lahko slovar potem opozori.

Uporaba korpusov je še posebej zanimiva za dinamična in z gospodarstvom neposredno povezana področja jezika, kot je terminologija. V korpusih lahko odkrijemo že uporabljene termine, njihove prevode ali razlage, s čimer je omogočeno bolj ažurno in cenejše izdelovanje slovarjev.

Ena prvih možnih uporab korpusa je za raznovrstne (formalne, socialne, literarne) jezikoslovne študije, predvsem za preverjanje teorij o jeziku skozi iskanje distribucije in konkretnih primerov izbranih pojavov. To velja toliko bolj za jezikovno ali kako drugače označene korpuse. V primerjavi z neobdelanim besedilom lahko v označenem korpusu iščemo bistveno bogatejše vzorce. Tako bi npr. za skladenjske raziskave bil zanimiv korpus, v katerem so besede oblikoslovno označene, za sociolingvistične pa npr. korpus, kjer je premi govor označen s spolom govorca.

Nenazadnje so računalniški korpusi pomembni za razvoj področja jezikovnih tehnologij, pa če so to pripomočki za avtorje, učenje jezikov ali prevajanje, programi za analizo in sintezo govora itd. Vsi takšni programi potrebujejo 'zavest' o jeziku, pri katerem naj bi bili v pomoč, potrebujejo torej računalniške jezikovne vire: slovarje, pravila in distribucije elementov določenega jezika. Mnogo teh virov je mogoče (pol)avtomatsko zajeti iz korpusov.

Standardi in označevanje korpusov

Računalniški korpusi besedil so dragoceni viri jezikovnih podatkov tako zaradi mnogoterih možnih uporab kot zaradi količine dela, ki ga je potrebno vložiti v njihovo izgradnjo. Ko to premoženje imamo, je smiselno omogočiti njegovo čim širšo uporabo (tj. *izmenljivost*) in ga zavarovati pred zastaranjem.

Na prvi pogled ravno računalniki zadovoljujejo ti dve želji, saj je razmnoževanje računalniških podatkov, za razliko od ostalih dobrin, praktično zastonj, digitalna informacija pa ne podleže zobu časa. Vendar morajo biti računalniški zapisi podrobno definirani, obenem pa so računalniki predmet bliskovitega tehnološkega razvoja. Zaradi tega se izkaže, da imajo besedila, hranjena na računalniških medijih, zaenkrat bistveno manjšo izmenljivost in trajnost kot pa tiskana besedila.

Problemi digitalnega zapisa besedil se začnejo že pri zapisu črk. Popolna računalniška podpora in soglasje o naborih znakov obstaja samo za angleško abecedo, medtem ko bomo v Sloveniji našli deset načinov, kako zapisati *č*, *š* in *ž*. Ker se vedno več besedil, ki sestavljajo korpus, zajema neposredno iz digitalnih virov, je problem različnih formatov dokumentov še posebej pereč; če se razlikujejo že zapisi črk, so toliko bolj različni načini zapisa odstavkov, premega govora, naslovov, opomb, bibliografskih podatkov itd. Razlikujejo se glede na programsko opremo, s katero je bilo besedilo narejeno, po videzu, kakršnega naj bi imelo tiskano besedilo, in glede na osebo, ki je besedilo napisala. Vendar so vsi ti podatki v korpusu vsaj potencialno pomembni, saj tvorijo del besedil, ki jih hočemo zajeti. Če v korpusu niso enotno in prepoznavno označeni, bo ta informacija izgubljena za uporabnike korpusa.

Četudi nam uspe pri izgradnji korpusa to zmedo na našem računalniku v lastno zadovoljstvo urediti, bodo na drugih računalnikih z drugimi operacijskimi sistemi in drugimi programi podatki vseeno neuporabni ali pa bodo vsaj zahtevali veliko truda za njihovo konverzijo v ciljni zapis. V primeru, da korpus še dodatno označimo (npr. s skladnjo, prevodi, leksikografskimi podatki), bo problem seveda še bistveno hujši. Izmenljivost takšnih zapisov je majhna.

Podobno majhna je tudi trajnost računalniških podatkov: besedila na petnajst let starem magnetnem traku so danes težko uporabna, podobno tudi besedila, napisana na urejevalniku teksta iz tistega časa. Ne enih ne drugih danes ne moremo več brati ali pa je v to potrebno vložiti precej truda.

Edino standardizacija lahko reši problem izmenljivosti in trajnosti digitaliziranih besedil. Poglavje v nadaljevanju obravnava tri nivoje tega procesa. Z osnovno in najbolj natančno definirano stopnjo računalniškega zapisa strukture besedil se ukvarja standard *SGML* (Standard Generalized Markup Language) mednarodne organizacije za standardizacijo ISO (International Organization for Standardization). Z zapisom in konkretnim označevanjem strukture besedil predvsem za namene znanstvene obravnave jezika se ukvarjajo s SGML skladna priporočila iniciative za označevanje besedil *TEI* (Text Encoding Initiative). Konkretno obliko zapisa računalniških korpusov za namene jezikovnih tehnologij pa podaja s TEI skladen zapis z imenom *CES* (Corpus Encoding Standard), ki nastaja oz. je nastajal v okviru evropske iniciative Eagles ter projektov MULTEXT in MULTEXT-East.

2.1 Standardni posplošeni jezik za označevanje

SGML (Standard Generalised Markup Language) [7] je ISO standard 8879, ki določa jezik za predstavitev dokumentov, nad katerimi bodo delovali programi za procesiranje besedil. Razlogi za izdelavo tega standarda so bili deloma omenjeni že zgoraj. V razvitih državah podjetja porabijo veliko časa in s tem denarja za iskanje in pripravo informacij, ki so pretežno besedila. Zato prihaja do potrebe po načinu zapisa, ki bo izmenljiv, odporen na tehnološke spremembe in ki bo omogočal uporabo dokumentov v različne namene. SGML je poskus takšnega zapisa.

SGML je prvenstveno jezik za *označevanje* dokumentov, pri čemer lahko oznake opisujejo kakršnokoli informacijo, ki je dodana osnovnemu besedilu, npr. podatek, da je neki niz v besedilu naslov, ime ali beseda, da je neka beseda glagol, da ima neki termin povezavo s svojo razlago, da neki stavek spremlja slika ali njegov prevod in da nek monolog govori Hamlet v prvem dejanju neke tragedije.

SGML se glede na ostale jezike za označevanje dokumentov odlikuje v treh karakteristikah:

*Poudarek na **opisnem** namesto postopkovnem označevanju.*

Za razliko od mnogih drugih formatov zapisa besedil (npr. Microsoftov RTF) so oznake SGML namenjene opisu lastnosti besedila, ki ga zajemajo, ne pa postopku, ki te lastnosti realizira na konkretnem mediju: oznaka npr. pove, da del besedila, ki ga zajema, predstavlja odstavek, ne pa, da je potrebno izpustiti prazno vrstico in za določeno mero zamakniti začetek naslednje vrstice. Opisno označeni podatki imajo to prednost, da vsebujejo informacije v bolj prečiščeni obliki in jih je zato lažje uporabiti v različne namene. Tako je en sam dokument (npr. slovar knjižnega jezika ali pa komplet tehničnih priročnikov) uporaben za izdajo v knjižni ali pa multimedialni CD-ROM obliki.

*Koncept **tipa** dokumenta.*

SGML bi lahko poimenovali tudi jezik za metaoznačevanje dokumentov, saj standard ne spregovori besede o tem, katere oznake moramo uporabljati in v kakšnih odnosih so te oznake med seboj. Namesto tega SGML vpelje pojem tipa dokumenta in z njim formalno *definicijo tipa dokumenta* DTD (Document Type Definition). Šele DTD konkretno določa, kako je lahko neki dokument strukturiran in kako izgledajo njegove oznake. Neki DTD tako predstavlja gramatiko za določen tip dokumentov, npr. za knjige, tabele, terminološke slovarje, scenarije itd. Takšen pristop omogoča široko aplikacijo standarda, saj tako lahko pokriva dokumente z izrazito različno strukturo. Verjetno je, vsaj posredno, najbolj znana definicija tipa dokumenta tista za HTML (Hypertext Markup Language), ki jo morajo upoštevati vse pravilno narejene strani svetovnega omrežja WWW (World Wide Web).

*Neodvisnosti od konkretnega **zapisa** besedil.*

Eden od osnovnih ciljev SGML je, da so v njem zapisani podatki, prenosljivi z ene strojne in programske opreme na drugo brez izgube informacije. SGML zato vsebuje splošen mehanizem za

nadomeščanje nizov ob procesiranju dokumenta. Z *entitetami* SGML je mogoče preseči neskladnosti in pomanjkljivosti v naborih znakov različnih specifičnih računalniških sistemov, saj lahko za neprenosljive znake definiramo opisna imena, tj. entitite.

*

Prostor tu ne dopušča obširnejše obravnave standarda SGML. Naj zadošča opomba, da v tujini vedno več podjetij, ki imajo opravka z velikimi količinami besedil (npr. proizvajalci opreme za svojo dokumentacijo, založniki, knjižnice itd.), prehaja na ta standard, obstaja pa tudi že kar nekaj podjetij, predvsem v ZDA in Zahodni Evropi, ki se ukvarjajo izključno s SGML, bodisi z izdelovanjem programske opreme ali pa, pogosteje, z omogočanjem končnim uporabnikom, da preidejo na ta standard. V Sloveniji zaenkrat še ni zaslediti aplikacij tega standarda ali pa njegove obravnave v literaturi; izjema je edino WWW stran V. Batagelja, ki podaja uvod v SGML [1].

2.2 Iniciativa za zapis besedil

TEI (Text Encoding Initiative) [11] se je začela na konferenci, ki je bila leta 1987 na Vassar Collegeu v New Yorku. Tam se je zbralo okoli trideset predstavnikov arhivarstva, znanstvenih ustanov ter raziskovalnih projektov, da bi obravnavali možnost izdelave standardnega zapisa besedil in da bi podali priporočila o njegovem obsegu, strukturi, vsebini in načinu izdelave. O zaželjenosti takšne pobude priča, da je TEI dobil podporo vseh najvplivnejših strokovnih združenj s področja računalniške obravnave besedil kot tudi s strani ameriške vlade in Evropske unije. TEI je prvi osnutek svojih priporočil (TEI P1) izdal leta 1990, drugega pa leta 1992. Medtem ko sta bila tako P1 kot P2 še osnutka, predstavlja leta 1994 izdan TEI P3 [15] zaključek prve faze dela TEI.

TEI je kot osnovo svojega zapisa vzel SGML. TEI P3 je nabor definicij tipov dokumentov in entitet, ki za široko paleto zvrsti besedil določa konkretne oznake in njihovo strukturo. Skorajda bolj pomembnih pa je 1200 strani dokumentacije, ki podaja pomen posameznih oznak, opisuje DTD-je ter izpelje način za njihovo kombiniranje ter nadgradnjo.

TEI P3 pozna tri vrste naborov oznak, ki jih sestavljamo v t. i. modelu Chicago pice. Vsaka pica ima dve nujni sestavini: paradižnik in sir. Podobno TEI loči *središčne oznake* (*core tags*), ki so obvezne v vseh s TEI skladnih dokumentih. Središčne oznake določajo nabore znakov, glavo dokumenta ter oznake, ki so na voljo v vseh TEI dokumentih, npr. oznake za naslove in odstavke.

Vsaka pica ima tudi testo kot osnovo, vendar se njegova zvrst (vsaj v Chicagu) lahko izbere: lahko je tanko in hrustljavo, lahko debelo in mehko, ne more pa biti oboje hkrati. Podobno se tudi besedila delijo na različne zvrsti, ki so med seboj razmeroma dobro ločene. *Osnovni nabori oznak* (*base tag sets*) v TEI P3 obsegajo osnovni nabor za leposlovje, poezijo, gledališče, zapis govora, tiskane slovarje ter terminološke baze.

Končno imajo pice lahko tudi enega ali več dodatkov, TEI pa *dodatne nabore oznak* (*additional tag sets*). Ti opisujejo raznovrstna dodatna označevanja, ki ponavadi predstavljajo določeno interpretacijo besedila ali pa netekstualne elemente besedil, kot so navzkrižne povezave (za stvarna kazala) ali pa slike. Takih naborov je vsega skupaj devet, med njimi so nabor za analitične mehanizme (npr. skladenjsko analizo), nabor za dokumentiranje uredniških posegov, nabor za imena in datume ter *nabor za jezikovne korpuse*.

Za konec poglejmo v sliki 1 še primera dveh delov dokumentov, ki sta zapisana v standardu SGML in skladno s priporočili TEI. Na levi je primer besedila, označenega s skladenjsko analizo, na desni pa del glave dokumenta, ki bi bila uporabna za zapis radijskih poročil. Bralec bo opazil, da so TEI oznake angleške: čeprav je v TEI obliki možno strukturirati zapis poljubnega jezika, ostaja metajezik zapisa angleški.

Vsi veliki korpusi, izdelani v zadnjih nekaj letih, so, če že ne dosledno sledili, vsaj upoštevali TEI priporočila, saj so le-ta najbolj podrobna in natančna določila za označevanje jezikovnih virov.

```
<p>
<cl type='finite declarative'
   function='independent'>
 <phr type=NP function='subject'>
   Nineteen fifty-four,
  <cl type='finite relative declarative'
   function='appositive'>
   when
   <phr type=NP function='subject'>
   I
   </phr>
   <phr type=VP function='predicate'>
    was eighteen years old
   </phr>
  </cl>
 ,
```

```
<sourceDesc>
 <scriptStmt id=CNN12>
 <bibl>
    <author>CNN Network News</>
    <title>News headlines</>
    <date>12 Jun 1991</>
 </bibl>
 </scriptStmt>
 <!-- this script statement
      might be used to document
      the parts of a spoken
      transcript which included
      a news broadcast -->
</sourceDesc>
```

Slika 1: Primera TEI označenih dokumentov.

2.3 Standard za zapis korpusov

Kljub temu da TEI P3 podaja tudi določila za zapis korpusov, so ta po eni strani za določene namene preveč kompleksna, po drugi pa v določenih podrobnostih tudi še pomanjkljiva. V okviru iniciative Evropske unije Eagles in evropskih projektov MULTEXT ter MULTEXT-East je v izdelavi SGML definicija tipa dokumenta z imenom CES (Corpus Encoding Standard) [10]. CES je v veliki meri skladen s priporočili TEI, vendar je enostavnejši in bolj ekspliciten, saj je njegova specifična domena opis večjezikovnih korpusov predvsem za namene jezikovnih tehnologij.

CES določa osnovni zapis in obseg označevanja, ki ga mora korpus zadovoljiti, da ga lahko še smatramo za standardiziranega. CES opredeli tri nivoje take standardizacije, kjer vsak višji nivo dodatno standardizira korpus:

1. CES-1 dokument ima s TEI skladno glavo, tj. bibliografske in ostale podatke o korpusu, telo dokumenta pa je označeno, v skladu s CES-definicijo dokumenta, z osnovno strukturo, tj. z glavnimi razdelki besedila do nivoja odstavkov.

2. CES-2 dokument ustreza nivoju CES-1, poleg tega pa vsebuje TEI oznake, na katere se lahko sklepa iz tipografskih informacij v originalnem besedilu: premi govor, imena, številke, datumi itd.

3. CES-3 dokument mora vsebovati CES-2 oznake, polega tega pa ustreza dodatnim zahtevam za označevanje stavkov in premega govora. Vse izključno tipografske informacije so odstranjene iz besedila in kvečjemu ohranjene kot vrednosti atributov.

4. Nivo jezikovnega označevanja: poleg osnovnih nivojev je korpus možno tudi dodatno označiti z jezikovnimi informacijami. CES obravnava dvoje takšnih označevanj, in sicer oblikoslovno označevanje besed in pa zapis poravnav v vzporednem korpusu, tj. zapis poravnave nekega elementa v originalu z njegovim prevodom.

3 Procesiranje korpusov

Poudarek prejšnjega poglavja je bil na standardih za opis jezikovnih podatkov. Seveda pa je programska oprema tista, ki nam omogoči neki korpus narediti oziroma izkoriščati. Ta proces lahko razdelimo na štiri dele, od katerih sta prva dva usmerjena v urejanje in eksplicitiranje jezikovne informacije, druga dva pa v njeno izkoriščanje. Tak pristop premakne težišče dela v izdelavo korpusa, s čimer olajša njegovo uporabo.

V izdelavi korpusa je (1) dobljena besedila najprej potrebno urediti in strukturno označiti, s čimer dobimo ekvivalent CES-2 oz. CES-3 standardiziranega korpusa. Korpus lahko (2) s pomočjo označevalcev še dodatno označimo z jezikovnimi podatki. Ta dva koraka zahtevata precejšen vložek človeškega dela, saj je podatke tu potrebno ročno vnašati ali pa vsaj preverjati. Vendar pa tako pridobimo dokumentiran in standardiziran jezikovni vir, ki je izmenljiv in ga lahko s široko dostopnimi orodji uporabljamo v raznovrstne namene.

Ker orodja za izkoriščanje korpusov dostikrat zahtevajo besedila v sebi lastnem formatu, je korpus (3) iz standardiziranega formata potrebno najprej pretvoriti v format orodja. Vendar pa je, za razliko od procesa izdelovanja korpusa, ta korak preprost in popolnoma avtomatski. Zadnji korak je seveda (4) dejanska uporaba korpusa s pomočjo ustreznih programov.

V nadaljevanju poglavja najprej obravnavamo orodja, ki imajo neposredno zvezo s SGML (1 in 3), nato (2) jezikovne označevalnike in končno (4) pregledovalnike korpusov.

3.1 Orodja SGML

Pri izgradnji korpusa je potrebno dobljena besedila najprej prevesti v SGML in TEI/CES. Besedila najprej očistimo podatkov, ki so odvečni namenu korpusa, in korpus označimo s podatki, ki so na enostaven način dosegljivi iz tipografskih vzorcev v besedilih. Ta pretvorba se tipično izvede z adhoc programi, napisanimi v katerem od programskih jezikov, ki je močan v iskanju in nadomeščanju vzorcev nad nizi. Za pogostejše formate zapisa besedil pa obstajajo tudi že napisani programi, ki dokument pretvorijo v osnovni zapis SGML. V tej fazi je tudi potrebno poskrbeti za glave posameznih besedil in celotnega korpusa, saj naj bi bil korpus (bibliografsko, pravno, uredniško ...) označen.

Ko je korpus vsaj v minimalnem zapisu SGML, je nad njim že mogoče uporabljati orodja SGML. Programsko opremo, ki se 'zaveda' standarda SGML, je mogoče kupiti, kar nekaj takšnih programov ali pa knjižnic pa je tudi prosto dostopnih. Osnovno orodje je razčlenjevalnik SGML, ki preveri, ali je neki dokument SGML v skladu s svojo definicijo tipa, in definira vsak element glede na njegovo mesto v tej definiciji. Ostali razredi programov SGML omogočajo enostaven vnos dokumetov SGML, iskanje podatkov v dokumentih ali pa pretvorbo iz zapisa SGML v ciljni zapis, npr. za tiskanje ali predstavitev na WWW, ali pa za prevedbo v neko specifično obliko, ki jo pozna naše orodje za pregledovanje korpusov.

Ker TEI zapis vsebuje več eksplicitne informacije kot pa originalni zapis, je v takšno konverzijo potrebno vložiti sorazmerno dosti dela: četudi so na voljo programi, ki bi konverzijo avtomatično opravljali, ti niso nezmotljivi, poleg tega pa prevod v rigorozni zapis SGML pogosto razkrije napake in nekonsistence v originalnih besedilih — takšne napake lahko bodisi popravimo bodisi označnimo kot napake.

Dodatno jezikovno označevanje se sicer tudi lahko že dogaja v SGML, vendar je poudarek pri tem že na jezikovnem znanju teh orodij, zato so obravnavana v naslednjem razdelku.

3.2 Jezikovno označevanje

Že za avtomatsko označevanje osnovnih jezikovnih informacij, npr. datumov ali pa stavkov, je potrebno nekaj znanja o jeziku; standardna oblika zapisa datumov se razlikuje od jezika do jezika, lahko pa je datum, ki ga hočemo kot takega označiti, napisan tudi z besedami, ali pa samo delno.

Za takšna označevanja se tipično uporablja adhoc programska oprema, čeprav so počasi že na voljo orodja, ki ta proces vsaj do določene mere parametrizirajo glede na jezik korpusa.

Kaj točno hočemo v besedilu označiti, je seveda odvisno od namembnosti korpusa. Tu bomo omenili dva programa, ki sta še posebej zanimiva za jezikoslovne in slovaropisne obravnave. Prvi oblikoslovno označi besede v besedilu, drugi pa stavčno poravna vzporedni korpus, oba pa spadata v razred programov, ki izkoriščajo statistične lastnosti jezika. Kot je bilo že omenjeno, so takšni programi v zadnjem času predmet velikega zanimanja [2], saj so robustni in se lahko učijo ob ročno označenih besedilih.

Za oblikoslovno označevanje besed v korpusu je potrebno najprej imeti slovar ali pa program, ki za besedne oblike določi njihove možne oblikoslovne označbe. Vendar pa ima neka besedna oblika tipično več možnih interpretacij: tako je npr. *berači* lahko glagol v velelniku ali povedniku, ali pa samostalnik v imenovalniku ali orodniku. V konkretnem besedilu pa bo besedna oblika imela seveda samo eno pravilno označbo. Naloga programov za oblikoslovno označevanje je, izmed možnih oblikoslovnih označb neke besede glede na sobesedilo določiti njeno pravo označbo.

Izdelanih je bilo že več označevalnikov, ki se lahko naučijo zakonitosti nekega jezika iz ročno označenih korpusov. Najbolj odmeven je bil verjetno t. i. označevalnik Xerox [3], ki z uporabo skritih označevalnih verig določi najbolj verjetno zaporedje oblikoslovnih označb besed v nekem stavku. Program ne zivaja skladenjske analize, pač pa izkorišča lokalni kontekst besede za določitev njene oznake. Za angleški jezik jezik doseže ta in njemu podobni označevalci približno 95 % natančnost. Za slovanske jezike je, kot kažejo preliminarni rezultati za češki jezik [8], ta natančnost verjetno manjša, in sicer približno 82 %.

Povsem drug način označevanja je mogoče uporabiti pri vzporednih korpusih. Tu je koristno določiti, kateri del originalnega besedila ustreza kateremu delu prevoda. Takšna paralelizacija je lahko bolj ali manj natanča: določimo lahko npr. samo povezave po poglavjih ali pa vse do povezav konkretnih besed v besedilu z njihovimi prevodi. Tudi tu je mogoče s statističnimi metodami doseči zadovoljive rezultate. Eden bolj zanimivih takšnih programov je opisan v [5]. Njegova odlika je predvsem enostavnost, saj samo iz števila znakov sklepa na najbolj verjetno povezavo med stavki originala in stavki prevoda. Tako z enostavnim orodjem dosežemo že precej koristen nivo paralelizacije.

Vsem programom za jezikovno označevanje je skupno, da je njihova točnost manj kot popolna. Za kakovosten korpus je zato koristno, da so dobljene oznake še ročno pregledane. Vendar pa to za velike korpuse postaja skorajda nemogoče, po drugi strani pa tudi ljudje ne označujejo popolnoma točno. Poleg enostavnih napak je problem tudi v tem, da vsako označevanje predstavlja interpretacijo besedila, ta pa se lahko od človeka do človeka razlikuje.

3.3 Pregledovalniki

Nad označenim korpusom lahko uporabimo raznovrstne programe, od katerih so najbolj zanimivi pregledovalniki. Ti morajo biti sposobni poiskati željene dele korpusa in informacijo ustrezno predstaviti. Najbolj znana oblika predstavitve informacij iz korpusa so konkordance ozirom t. i. prikaz KWIC (key-word in context). Tu so pojavitve izbrane besede ali sobesedja poravnano izpisane skupaj s svojim sobesedilom.

Kot primer konkordanc je v sliki 2 podano nekaj pojavitev besede *mulatjera*; korpus, iz katerega je bila ta konkordanca narejena, je računalniška konferenca GORE iz omrežja SLON. Kot zanimivost še povejmo, da te besede ne najdemo niti v Verbinčevem Slovarju tujk niti v Slovarju slovenskega knjižnega jezika.

```
žnejša pobočja , kjer je speljana <mulatjera> . Po njih brez težav rahlo p
ja so skorajda brez snega , le po <mulatjeri> moraš večino časa gaziti do
od Kanceljni proti lovski koči ob <mulatjeri> na Kriške pode . Po treh ura
iguje proti Vratom . Na Vrata nas <mulatjera> pripelje v dveh urah . V vro
l na južno stran in jo uberemo po <mulatjeri> strmo vsekani navzgor v vršn
e . To je bila očitno oskrbovalna <mulatjera> in je zato speljana ves čas
 zelo blizu vršne piramide Krna . <Mulatjera> se tu neopazno zgubi . Na se
ne Zaprikraj . Pot je speljana po <mulatjerah> iz prve svetovne vojne mimo
a lovska stezica okoli Pihavca do <mulatjere> , ki pelje iz Trente na Kriš
 ( Robon ) in povratkom po daljši <mulatjeri> čez pode na Nevejsko sedlo .
izrazite doline pod Hudim vrhom . <Mulatjera> na Lipnik je namreč speljana
m je orientacijskih težav konec . <Mulatjera> nas v lepih ključih pripelje
asel nekaj , kar je spominjalo na <mulatjero> . Bilo je se vec kot meter s
Tista petstoenka je prav uživaška <mulatjera> . Prvo uro te vodi po vršnem
edvsem biti vsaj malo mazohista . <Mulatjera> pa je izredno prijetna in pr
isi po stari vojaški cesti ( bolj <mulatjeri> ) skozi tunel , ali pa prav
 je še dovolj snega , zato sva po <mulatjeri> smučala . Nekje na tretjini
 Čez 15 min se na levem ovinku te <mulatjere> napotiš kar naravnost čez go
```

Slika 2: Primer formata KWIC.

Bolj kot iskanje posameznih besed je zanimivo iskanje sobesedij. Te, ti. *kolokacije*, namreč lahko razkrijejo vezave besed tako s skladenjskega kot s pomenskega stališča. Način pregledovanja je podoben kot pri KWIC, obstajajo pa tudi programi, ki avtomatsko izberejo sopojavitve, ki so statistično in zato verjetno tudi jezikovno signifikantne. Možnost iskanja kolokacij je toliko bolj zanimiva za označene korpuse, saj tu lahko iščemo tudi sopojavitve bolj abstraktnih kategorij.

Končno je tu še paralelno prikazovanje vzporednih korpusov. Prikaz je tipično v dveh poravnanih (KWIC-) oknih, iskalni jezik orodij, ki takšne korpuse podpirajo, pa razširjen tako, da se lahko kriteriji za iskanje nanašajo na več vzporednih besedil. Natančneje ko so korpusi povezani, bolj podrobno je lahko takšno iskanje. Zanimivo je, da so vzporedni korpusi primerni tudi za enojezikovne raziskave. Tako npr. iskanje vseh pojavitev neke besede, katere prevod se ne pojavi v prevodu stavka, v katerem se beseda nahaja, hitro pokaže na idiomatske uprabe te besede.

Programe za prikazovanje korpusov je možno kupiti, nekateri so pa tudi prosto dostopni. Vendar je zagotovitev ustreznega pregledovalnika še vedno problematična, saj vsi ne tečejo na vseh računalniških platformah, imajo nepopolno funkcionalnost ali pa ne delujejo pravilno za slovenski jezik. Glede na veliko razširjenost TEI za zapis korpusov se v zadnjem času posebno pozornost posveča pregledovalnikom, ki delujejo nad označenimi korpusi SGML. Takšni pregledovalniki imajo prednost, da lahko izkoristijo vse oznake (npr. bibliografske) in da so v precej večji meri jezikovno neodvisni.

4 MULTEXT-East

V Skupini za govor in jezik Odseka za inteligentne sisteme na IJS sodelujemo v evropskem projektu MULTEXT-East (Multilingual Text Tools and Corpora for Central and Eastern European Languages, [4]). Projekt je podaljšek evropskega projekta MULTEXT, v katerem so sodelovale inštitucije iz šestih držav članic Evropske unije. MULTEXT-East je dvoletni projekt, ki se je začel maja 1996, v njem pa sodeluje poleg koordinatorja iz Aix-en-Provance in pridruženega partnerja iz Pise še šest skupin iz držav srednje in vzhodne Evrope, in sicer Bolgarije, Češke, Estonije, Madžarske, Romunije in Slovenije.

4.1 Korpus

Eden od ciljev MULTEXT-East je proizvesti standardiziran večjezikovni korpus, ki vsebuje približno dva milijona besed, sestavljen pa je iz naslednjih delov:

1. *vzporedni korpus*, ki vsebuje roman 1984 G. Orwella v originalu in prevode v šestih jezikih projekta (približno 7 × 100.000 besed);

2. *primerljiv korpus*, sestavljen iz dveh nadaljnjih delov: prvi vsebuje šest leposlovnih del avtorjev iz držav članic projekta, drugi pa šest zbirk časopisnih člankov v jezikih teh držav (približno 2 × 6 × 100.000 besed); slovenski del primerljivega korpusa je sestavljen iz romana Galjot D. Jančarja in 45 člankov iz časopisa Dnevnik;

3. *govorjeni korpus*, sestavljen iz 40 krajših odlomkov iz evropskega projekta EUROM, prevedenih v šest jezikov projekta (približno 7 × 2.500 besed), prebranih in digitaliziranih, pri čemer bo ta govorjeni korpus poravnan s svojim ortografskim zapisom.

Celoten korpus je označen po priporočilu CES; poleg bibliografskih bodo oznake vsebovale strukturne informacije (odstavki, članki, naslovi, premi govor, itd.) ter določene »posebne besede«, npr. lastna imena in okrajšave. Kot primer, kako takšne označbe izgledajo, sta v sliki 3 podana dva odlomka iz slovenskega in češkega prevoda romana 1984.

svojem sedežu, njegove mogočne prsi so se napenjale in trzale, kakor bi se upirale naskoku valov. Temnolaso dekle za	by spolkl pravítko a jeho mohutná hruď se nadouvala a zachvívala, jako by čelil náporu vln. Tmavovlasá dívka za
`<name type=person>`Winstonom`</name>`	`<name type=person>`Winstonem`</name>`
je začelo kričati	začala vykřikovat
`<q>`	`<q rend='PRE lsquo POST rsquo'>`
Svinja! Svinja! Svinja!	Svině!
`</q>`	`</q>`
in nenadoma je pograbila težak slovar	a zničehonic popadla těžký slovník
`<foreign lang=ns>`Novoreka`</foreign>`	`<foreign lang=ns>`newspeaku`</foreign>`

Slika 3: Slovenska in češka odlomka iz »1984«.

Del korpusa bo tudi dodatno označen: vsi prevodi 1984 bodo stavčno poravnani z originalom, medtem ko bo del korpusa označen še z oblikoslovnimi oznakami.

4.2 Oblikoslovje: definicija, slovar, označevanje

Oblikoslovno označevanje je najzahtevnejši del nadgradnje osnovnega (CES-3) korpusa. Da lahko (pol)avtomatsko označimo besedne oblike v korpusu z njihovimi oblikoslovnimi oznakami, so potrebni naslednji koraki: definirati je potrebno oblikoslovne kategorije, nato izdelati slovar, ki za vsako besedno obliko določi njene možne oznake, sestavljene iz oblikoslovnih kategorij, in z njegovo pomočjo polavtomatsko označiti besedila.

```
Verb (V)

14 Positions

**** **** **** **** **** **** **** ---- ---- ---- ---- ---- ---- ----
PoS  Type VFrm Tens Pers Numb Gend Voic Neg  Def  Cltc Case Anim Clt2
****' **** **** **** **** **** **** ---- ---- ---- ---- ---- ---- ----

= ============== ============== =  RO  SL  CS  BG  ET  HU
P ATT            VAL            C  x   x   x   x   x   x
= ============== ============== =
1 Type           main           m  x   x   x   x   x   x
                 auxiliary      a  x   x   x   x   x   x
                 modal          o  x   x   x       x
                 copula         c  x   x   x
- -------------- -------------- -
2 VForm          indicative     i  x   x   x   x   x   x
                 subjunctive    s  x
                 imperative     m  x   x   x   x   x   x
                 conditional    c      x   x       x   x
                 infinitive     n  x   x   x       x   x
                 participle     p  x   x   x   x   x
                 gerund         g  x       x   x   x
                 supine         u      x       x
               l.s. transgressive t              x
               l.s. quotative     q              x
- -------------- -------------- -
...
```

Slika 4: Začetek MULTEXT-*East tabele za glagol.*

Ker je projekt večjezikoven, je potrebno oblikoslovne oznake definirati v skupnem formatu za šest jezikov. Kot primer iz MULTEXT-East 'slovnice' je v sliki 4 podan začetek tabele za glagole: ta določa, da glagolsko besedo opisuje 14 lastnosti. Najprej je podana besedna vrsta (tj. glagol, V), v tabeli pa vidimo definicijo prvih dveh lastnosti glagola; za vsako lastnost je podano ime ter nabor njenih vrednosti. Imenu vrednosti sledi njena enočrkovna koda le-te ter določitev, katere jezike opisuje. Tako npr. slovenščina loči glagolske oblike povednika, velelnika, pogojnika, nedoločnika, deležnika ter namenilnika.

Že iz zgornjega bo jasno, da določitve MULTEXT-East za oblikoslovje mestoma odstopajo od tradicionalnih kategorij v slovenskih slovnicah; tako so npr. glagolska deležja in glagolniki razvrščeni med prislove in samostalnike. Takšna odstopanja so v veliki meri posledica usklajevanja zapisov šestih med seboj zelo različnih jezikov, posredno pa dvanajstih, saj so tabele usklajene tudi z jeziki MULTEXT.

Predstavljeni format ima to prednost, da je neko oblikoslovno oznako mogoče zapisati v kompaktnem, obenem pa še vedno berljivem (ASCII) zapisu: tako npr. niz Vmip3s določa

vrednosti Verb main indicative present third singular oz. povednik glavnega glagola v tretji osebi ednine.

Naslednji korak je izdelava slovarjev, ki bodo v MULTEXT-East vsebovali 15.000 gesel za vsakega od šestih jezikov projekta. Ti slovarji poleg samih korpusov predstavljajo tudi pomemben vir jezikovnih podatkov.

```
berači   berač      Ncmpi
berači   berač      Ncmpn
berači   beračiti   Vmip3s--n
berači   beračiti   Vmmp2s
```

Slika 5: Fragment MULTEXT-*East slovarja.*

Slovarji imajo preprosto, pa vendar precej informativno strukturo: vsak vnos je sestavljen iz besedne oblike, njenega gesla ter njenih oblikoslovnih značilnosti. Primer vnosov za besedno obliko *berači* je podan v sliki 5.

S slovarjem je nato mogoče začeti označevanje besed v korpusu. Glavni problem takšnega označevanja je seveda dvoumnost besednih oblik — tako ima *berači* štiri možne interpretacije, od katerih bo na določenen mestu v besedilu samo ena pravilna.

Kot je bilo že rečeno, je za avtomatsko določanje pravilne oznake mogoče uporabiti statistične označevalnike, vendar pa ti potrebujejo ročno označen korpus za učenje. Ker tak korpus za slovenski jezik (pa tudi za ostale jezike projekta, razen češkega) ne obstaja, bo v okviru projekta potrebno ročno označiti del korpusa, nato pa v zaporedju več korakov izšolati označevalec, ročno popraviti rezultate in postopek nato ponoviti na razširjeni učni množici. Ker označevalci potrebujejo velike učne množice, ročno pregledovanje pa je izredno zamudno delo, bodo rezultati projekta tu samo pripravljalni. Verjetno bo ročno pregledan samo del korpusa, ker pa je potrebna velikost učne množice odvisna tudi od števila možnih oznak, bo število oblikoslovnih oznak v besedilu zgoščeno glede števila slovarskih oznak.

4.3 Dostop do rezultatov projekta

Kot je bilo že rečeno, projekt še teče, vendar je precejšnje število vmesnih rezultatov že dostopno. Ker uporaba zgrajenih virov pokaže na napake in pomanjkljivosti teh virov, bodo rezultati dostopni v dokončni obliki šele ob koncu projekta, vmesni rezultati pa obsegajo zbran, dokumentiran in bibliografsko ter strukturno označen korpus, definirane oblikoslovne tabele in prvo verzijo slovarja. V nadaljevanju projekta je potrebno izdelati še končne verzije teh virov, stavčno paralelizirati vzporedni korpus ter korpus oblikoslovno označiti.

S tem bo izdelanih nekaj osnovnih računalniških virov za slovenski jezik, ki bodo usklajeni z mednarodnimi standardi in priporočili ter s petimi drugimi jeziki projekta. Kljub temu da so ti viri premajhni za marsikatero aplikacijo, so vendarle pomembni, saj bodo prvi tovrstni široko dostopni viri slovenskega jezika — rezultati projekta bodo namreč v neprofitne namene dostopni zastonj. Vsaj za našo skupino na IJS pa so verjetno bolj kot izdelava samih virov pomembne izkušnje, ki smo jih pridobili pri projektu, saj predstavljajo osnovo, na kateri bi bilo mogoče zgraditi referenčni korpus slovenskega jezika.

Za popularizacijo (rezultatov) projekta smo na IJS postavili WWW stran z naslovom http://nl.ijs.si/ME, ki vsebuje vse osnovne informacije o projektu, primere iz korpusa ter slovarjev, pa tudi vmesne rezultate projekta.

5 Zaključek

Članek je predstavil nekatere vidike računalniških zbirk besedil. Uporabnost takšnih korpusov je nedvomna, vendar pa je njihova izdelava, razširjanje in uporaba razmeroma zahtevna. Članek se je osredotočil na bolj tehnične vidike zapisa, izdelave in uporabnosti korpusov, izpustil pa je obravnavo ravno tako pomembnih pravnih in človeških vidikov — teh se na kratko dotaknemo tu.

Pravno vprašanje je izredno pomembno, posebno pri izdelavi, saj so besedila v korpusu še vedno last avtorjev, založb ali prevajalcev. Ti se ponavadi bolj ali manj upravičeno bojijo zaupati svoja besedila na računalniškemu mediju urednikom korpusa in nato tretjim osebam, saj je, vsaj v teoriji, ta besedila nato možno razmeroma enostavno ponatisniti ali kako drugače neavtorizirano uporabiti. Izkušnje evropskih projektov kažejo, da uredniki korpusov porabijo ponavadi skorajda več časa za pridobitev privoljenj lastnikov besedil kot pa nato za izdelavo samega korpusa. Pravni status dodatno zapleta dejstvo, da korpus sestavljajo tudi oznake v njem, te pa so last urednikov korpusa.

Če so uredniki korpusa tudi njegovi edini uporabniki, je neavtorizirana uporaba še relativno enostavno obvladljiva. Vendar pa utegne korpus biti zanimiv tudi tretjim osebam. Ob predpostavki, da lastniki besedil, pa tudi uredniki korpusov, zaupajo pravnemu sistemu svoje države, je možno nezaželjeno izkoriščanje korpusov urediti z ustreznimi izjavami, s katerimi se morajo zavezati tako uredniki korpusov kot tudi nadaljnji uporabniki. Formuliranje takšnih izjav na srečo postaja vedno lažje, saj je na voljo že dosti primerov iz evropskih projektov.

Seveda pa je izdelava korpusov, posebno široko dostopnih, smiselna samo, če se ti korpusi nato tudi uporabljajo. Tu stopi v ospredje človeški dejavnik, saj dosti institucij, ki bi takšne korpuse lahko s pridom uporabljale, nima razvite računalniške ekspertize. Verjetno je najlažji način, kako takšni jezikovni viri lahko zaživijo, uvajanje njihove uporabe v primerne visokošolske študije ter s popularizacijo celotnega področja jezikovnih tehnologij.

Kot je bilo že rečeno, javno dostopnih in standardiziranih korpusov za slovenski jezik še ni. Izdelava korpusov in ostalih jezikovnih virov je predraga, da bi bilo smiselno že v prvi fazi prepustiti njihov nastanek ekonomskim dejavnikom, še posebej za jezike s tako majhnim številom govorcev, kot jih ima slovenski jezik. Z vladnim financiranjem in sodelovanjem založb, računalniških hiš in akademskih institucij bi bilo nujno najprej omogočiti izdelavo široko dostopnih virov, saj šele ti lahko dajo eno od prepotrebnih osnov za nadaljnji razvoj raziskovanja in uporabe naše materinščine.

Literatura

[1] Vladimir Batagelj (1995). Uvod v SGML. URL: http://vlado.mat.uni-lj.si/vlado/sgml/sgmluvod.htm.

[2] Eugene Charniak (1994). *Statistical Language Learning*. Language in Computers 12. The MIT Press.

[3] D. Cutting, J. Kupiec, J. Pedersen in P. Sibun (1992). A Practical Part-of-Speech Tagger. V: *Proceedings of the Third Conference on Applied Natural Language Processing*, str. 133-140, Trento, Italija.

[4] Tomaž Erjavec, Nancy Ide, Vladimír Petkevič in Jean Véronis (1996). Multext-east: Multilingual Text Tools and Corpora for Central and Eastern European Languages. V: *Proceedings of the First TELRI European Seminar: Language Resources for Language Technology*, str. 87-98.

[5] William Gale in Ken W. Church (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19 (1): 75-102.

[6] R. Garside, G. L. Leech in G. Sampson (uredniki) (1987). *The Computational Analysis of English*. London in New York: Longman.

[7] Charles F. Goldfarb (1990). *The SGML Handbook*. Oxford: Clarendon Press.

[8] Barbora Hladka in Jan Hajič (1996). Tagging a Highly Inflected Language. V: *Proceedings of the First TELRI European Seminar: Language Resources for Language Technology*, 191-196.

[9] Miran Hladnik (urednik). Zbirka slovenskih leposlovnih besedil. URL: http://www.ijs.si/lit/leposl.html.

[10] Nancy Ide, Greg Priest-Dorman in Jean Véronis (1996). Corpus Encoding Standard V1.3. Tehnično poročilo, Eagles, Multext & Multext-East, http://www.cs.vassar.edu/CES/CES1.html.

[11] Nancy Ide in Jean Véronis (urednika) (1995). *The Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic Publishers.

[12] Henry Kučera in William Nelson Francis (1967). *Computational Analysis of Present Day American English*. Rhode Island: Providence, Brown University Press.

[13] John Sinclair (urednik) (1987). *Looking Up: An account of the COBUILD Project in lexical computing*. Collins.

[14] John Sinclair (1994). Corpus Typology. EAGLES DOCUMENT EAG–CSG/IR–T1.1, Commission of the European Communities.

[15] C. M. Sperberg-McQueen in Lou Burnard (urednika) (1994). *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford.

[16] Jože Toporišič (urednik) (1975). *Besedila slovenskega jezika*. Ljubljana: Filozofska fakulteta.

Tomaž Erjavec

UDK 81:681.3

SUMMARY

COMPUTERIZED TEXT COLLECTIONS

Ordered and computerized text collections — corpora — are becoming an indispensable source of linguistic data. Freely available corpora of the Slovene language do not yet exist. The article gives a historical overview of the development of computer corpora, their typology and fields of application. Two aspects of corpora are discussed next: the standardization of their encoding and the tools for their development and exploitation. The second part of the article gives an overview of the MULTEXT-East project (Multilingual TextTools and Corpora for Central and Eastern European Languages), which also includes the Slovene language. The focus of the presentation is on the corpus and morphosyntactic descriptions developed in the project and on its currently available results. Finally, some possibilities for developing the field of corpus linguistics in Slovenia are discussed.