SLOVENSKO-ANGLE[KI KORPUS *ELAN*

Ve~jezi~ni korpusi so pomemben vir podatkov prevodoslovnim raziskavam in tehnologijam strojnega in strojno podprtega prevajanja. ^lanek predstavi dvojezi~ni korpus, ki smo ga na IJS zbrali v okviru projekta EU ELAN. Korpus vsebuje milijon besed, sestavlja pa ga 15 sodobnih terminolo{ko bogatih besedil in njihovih prevodov v slovenskem in angle{kem jeziku. Besedila so stav~no poravnana ter ozna~ena v skladu s predpisi TEI (Guidelines for Text Encoding and Interchange). Korpus in vsako besedilo v njem je ozna~eno z glavo TEI, poravnana dvobesedila pa so shranjena podobno kot pomnilniki prevodov. Celoten korpus je dostopen na mre‘nem naslovu http://nl.ijs.si/elan/, kjer so na voljo tudi dvojezi~ne konkordance korpusa.

Uvod

Pomen ve~jezi~nih korpusov se ka‘e skozi veliko {tevilo projektov, ki se jih trudijo zagotoviti: med evropskimi so zgled MLCC (Armstrong et al. 1998) (devet jezikov EU), Crater (McEnery et al. 1997) ({pansko-francosko-angle{ko) in ENPC (Johansson et al. 1996) (angle{ko-norve{ko). Za slovenski jezik je bil doslej edini ve~jezi~ni korpus objavljen na CD-ROM-u v okviru evropskega zdru‘enja TELRI (Erjavec et al. 1998); vzporedni korpus na tej zgo{~enki vsebuje Platonovo Republiko, delo In{tituta za slovenski jezik pri ZRC SAZU, in korpus projekta MULTEXT-East (Erjavec, Ide 1998). ^eprav CD-ROM TELRI ponuja veliko {tevilo jezikov, je zapisan v skladu z mednarodnimi priporo~ili in bogato ozna~en, pa vsebuje le malo slovenskih besedil (pribl. 500.000 besed iz {tirih besedil), ki hkrati niso najbolj primerna za terminolo{ke raziskave, torej za podro~je, na katerem so vzporedni korpusi {e najhitreje uporabni.

S projektom EU ELAN (the European Language Activity Network) se je ponudila prilo‘nost, da se izbolj{a ponudba korpusov za slovenski jezik. V okviru projekta smo zbrali, ozna~ili in uredili korpus, ki vsebuje milijon besed v 32.000 prevodnih enotah, od katerih vsaka vsebuje segment (ve~inoma poved) v slovenskem in angle{kem jeziku. Skladno z nameni projekta je korpus prosto dostopen tako za prepis na ra~unalnik kot za iskanje. ^lanek predstavi korpus IJS-ELAN in ima naslednjo zgradbo: poglavje 2 opi{e na~in zbiranja in obdelave besedil, poglavje 3 na{teje 15 besedil, zbranih v korpusu, poglavje 4 poda ra~unalni{ki zapis (oznake) korpusa, poglavje 5 mo‘nosti dostopa do korpusa in poglavje 6 zaklju~ke.

Zagotovitev besedil in obdelava

Ker smo imeli za izdelavo korpusa IJS-ELAN na voljo samo pol leta, je bil eden od osnovnih ciljev zbrati ~im ve~ji korpus s tem, da zaobidemo najbolj zamudne korake v izdelavi korpusa. Proces izdelave (urejanja), ki je podrobneje opisan v (Erjavec 1999b), je pri korpusu IJS-ELAN zajemal: pridobitev in za{~ito avtorskih pravic do izvirnih besedil in korpusom kot celoto, zagotovitev digitalnih izvirnikov besedil, segmentacijo in poravnavo povedi, raz~lenitev na besede (tokenizacija), zapis v standardizirani format; zapis glav korpusa in besedil in zagotovitev dostopnosti na svetovnem spletu (WWW). Klju~nega pomena za uspe{nost projekta so bili na~in pridobivanja besedil, izdelave poravnanih dvobesedil ter njihov ra~unalni{ki zapis.

Sodelavci projekta so besedila zbirali sami ter jih nato tudi pretvorili iz izvirnega zapisa, jih segmentirali in poravnali. Ker je korpus namenjen ~im bolj svobodnemu raz{ir-

janju, smo izbrali besedila, kjer avtorske pravice niso problemati~ne, bodisi ker besedila sodijo med javne vladne dokumente bodisi med dokumentacijo programja GNU, kjer je nadaljne raz{irjanje celo za‘eleno, za nekatera besedila pa so bile urejene ‘e prej v okviru drugih projektov. Reprodukcija besedil v celoti na osnovi korpusa (kar je po navadi glavna ovira nadaljnjemu raz{irjanju) pa je ote‘ena z na~inom zapisa korpusa.

Na~in pretvorbe in poravnave se je med sodelavci razlikoval; uporabili smo orodja UNIX in pomnilnik prevodov Deja Vu podjetja Atril. V obeh primerih smo kot rezultat dobili poravnana dvobesedila, ki so v precej{nji meri o~i{~ena izvirnih oznak (npr. HTML, RTF) in zapisana v enostavnem tabelari~nem formatu, po ena prevodna enota (tj. poved v izvirniku in njen prevod) v vsaki vrstici. ^eprav prevodne enote v splo{nem ustrezajo povedim, so v~asih tudi dalj{e ali pa kraj{e, saj se zgodi, da eni povedi v izvirniku ustrezata dve ali ve~ v prevodu ali obratno.

Tako dobljena dvobesedila smo nato o~istili s filtri, napisanimi v programskem jeziku Perl, s ~imer smo normalizirali zapis nabora znakov in odstranili ostanke formatiranja. Naslednji korak je bila tokenizacija, tj. identifikacija besed in lo~il, za kar smo uporabili orodje MULTEXT ’mtseg’ (Di Cristo 1996) s pravili MULTEXT-East za slovenski in angle{ki jezik (Dimitrova et al. 1998). Tokenizacija ozna~i tudi {tevila, okraj{ave itd. Tudi ta korak zaradi nepopolnih orodij in pravil v korpus prinese napake, ki so spet prete‘no popravljene s filtri Perl. Tokenizirana poravnana dvobesedila smo nato pretvorili v standardizirano obliko TEI, kjer se jim je tudi dodal opis v obliki glav TEI. Zadnji korak je bil pakiranje korpusa za prenos preko mre‘e ter pretvorba iz standardne oblike v tak{no, ki je primerna za WWW in slu‘i kot vir informacij o korpusu, ter v zapis za konkordan~nik, ki je ravno tako priklopljen na WWW. Tu smo uporabili prosto dostopen program Omnimark Lite, ki zna ravnati z dokumenti, zapisanimi v skladu s standardom SGML.

Zvrstnost korpusa

[tevilo, velikost in vrste besedil v korpusu je kompromis med (pri~akovano) uporabnostjo in enostavnostjo zagotovitve pravic in digitalnega originala. K uporabnosti prispeva predvsem dejstvo, da gre za sodobna besedila (90. leta), ki so terminolo{ko bogata in z zanimivih in dinami~nih podro~ij. Zaradi enostavnosti smo ve~inoma vklju~ili besedila, ki nimajo posebnih omejitev nad nadaljnjim raz{irjanjem in so nam bila dostopna v digitalni obliki na WWW. S Slu‘bo Vlade RS za evropske zadeve, kjer so nam prepustili ve~jo koli~ino {e ne objavljenih besedil, pa smo podpisali posebno pogodbo.

Korpus je sestavljen iz petnajstih enot, ki so ve~inoma integralna dvobesedila, vendar brez prete‘no nebesedilnih delov (npr. tabel s {tevilkami). Vsako dvobesedilo ima pripisan identifikacijski niz in skupaj s svojo glavo predstavlja element korpusa. Dvobesedila so razdeljena na tista s slovenskim izvirnikom in angle{ki prevodom, in tista z angle{kim izvirnikom in prevodom v slovenski jezik. Poleg razlike po smeri prevoda imata ta dva dela tudi precej razli~no zvrstnost.

Polovica s slovenskim izvirnikom je ve~inoma vladnega izvora, sestavlja pa jo enajst enot. Te enote skupaj s svojim identifikatorjem, pribli‘no velikostjo v kB in {tevilom besed v tiso~ih, letnico izida, kratkim naslovom ter zalo‘bo so naslednje:

usta: 364 Kb, 20 kW, 1997
Ustava Republike Slovenije; Ustavno sodi{~e Republike Slovenije

kuca: 1102 Kb, 69 kW, 1990-95
Govori predsednika RS, M. Ku~ana; Urad predsednika Republike Slovenije

parl: 325 Kb, 20 kW, 1998
Delovanje Dr‘avnega zbora; Dr‘avni zbor Republike Slovenije

ecmr: 4056 Kb, 239 kW, 1998/1999
Ekonomsko ogledalo; 13 {tevilk; Urad Republike Slovenije za makroekonomske analize in razvoj

ekol: 1222 Kb, 70 kW, 1999
Nacionalni program varstva okolja; Republika Slovenija, Ministrstvo za okolje in prostor, Uprava RS za varstvo narave

spor: 589 Kb, 34 kW, 1996
Evropski sporazum; Slu‘ba Vlade RS za evropske zadeve

anx2: 483 Kb, 25 kW, 1996
Evropski sporazum – Priloga II; Slu‘ba Vlade RS za evropske zadeve

stra: 1511 Kb, 89 kW, 1997
Strategija Slovenije za vklju~evanje v EU; Slu‘ba Vlade RS za evropske zadeve

kmet: 543 Kb, 29 kW
Dr‘avni program za prilagajanje zakonodaje – kmetijstvo; Slu‘ba Vlade RS za evropske zadeve

ekon: 394 Kb, 23 kW
Dr‘avni program za prilagajanje zakonodaje – gospodarstvo; Slu‘ba Vlade RS za evropske zadeve

vade: 471 Kb, 24 kW, 1995
Vademecum Lekove doma~e lekarne; Lek d.d.; OTC Division

^eprav del z angle{kim izvirnikom vsebuje skoraj polovico besed v korpusu, je sestavljen iz samo {tirih enot, od katerih sta dve knjigi. Vsebuje tudi druga~ne besedilne vrste in podro~ja kot prva polovica: dve enoti se ukvarjata z ra~unalni{tvom, ena pa z vizijo totalitarne dru‘be:

vino: 1182 Kb, 69 kW, 1994
EC Council Regulation No 3290/94 – agriculture / Uredba sveta ES {t. 3290/94 – kmetijstvo; Slu‘ba Vlade RS za evropske zadeve

ligs: 3044 Kb, 173 kW, 1999
Linux Installation and Getting Started / Namestitev in za~etek dela z Linuxom; Linux Documentation Project (-en: Specialized Systems Consultants; -sl: Linux User Group of Slovenia, LUGOS)

gnpo: 353 Kb, 13 kW, 1999
GNU PO localisation files / GNU PO lokalizacije datoteke; Free Software Foundation, Linux Documentation Project

orwl: 6698 Kb, 195 kW, 1983
G. Orwell: Nineteen Eighty-Four / 1984; projekt MULTEXT-East, slovenski prevod: knji‘nica Kondor, Mladinska knjiga (prevajalka: Alenka Puhar).

Kot je razvidno iz seznama, vsebuje korpus raznovrstna, ter ve~inoma terminolo{ko bogata besedila, primerna za avtomatsko identifikacijo terminov in njihovih prevodov. Za~etne raziskave (Vintar, 1999) z uporabo prete‘no statisti~nih metod ka‘ejo, da je korpus mo‘no uporabiti v terminolo{ke namene.

Zapis korpusa

Korpus je zapisan v skladu standardom SGML (Standard Generalized Markup Language, ISO 8879) in uporablja definicijo tipa dokumentov, ki je parametrizacija priporo~il TEI (Sperberg-McQueen, Burnard 1994). ^eprav vsebujejo priporo~ila TEI tudi predlog za zapis vzporednih poravnanih korpusov, se nam ta niso zdela primerna za na{ korpus. Namesto tega smo parametrizirali TEI tako, da je zapis bolj podoben tistemu, ki se uporablja pri pomnilnikih prevodov (Erjavec 1999a) .

V na{em zapisu uporabljamo generi~ne elemente TEI za zapis glav ter za oznake znotraj segmentov (povedi), medtem ko zapis besedil spremenimo tako, da ta neposredno vsebujejo prevodne enote, tj. so poravnana dvobesedila. Nabor znakov v korpusu je definiran opisno, z mehanizmom entitet SGML. Tako je npr. ~ v korpusu zapisan kot `&ccaron;`,] kot `&Cacute;`, & pa kot `&`, definicija tipa korpusa pa vsebuje nabor in opis uporabljenih entitet. Celoten korpus je sestavljen iz glave korpusa, ki vsebuje informacije o korpusu kot celoti in iz petnajstih elementov korpusa; vsak od teh spet vsebuje glavo in telo (dvobesedilo).

Glava TEI vsebuje podatke o datoteki, o viru ali virih besedila, o zapisu besedila in seznam sprememb. ^eprav je jezik projekta, v okviru katerega je korpus nastajal, angle{ki, smo poleg angle{~ine v glavah uporabljali tudi slovenski jezik. Za vtis o tem, kak{ne podatke vsebujejo glave v korpusu, podamo nekaj primerov. Prvi je za~etek korpusa in njegove glave:

```
<teiCorpus.2>
  <teiheader type="corpus" lang="slen" id="ijs-elan" creator="et"
  status="update" date.created="1999-04-14" date.updated="1999-06-22">
  <filedesc>
   <titlestmt>
    <title lang="en">The IJS-ELAN Slovene/English Parallel Corpus</title>
     <title lang="sl">Slovenskoangle&scaron;ki vzporedni korpus IJS-ELAN</title>
```

Naslednji primer iz glave korpusa poda deklaracijo oznak, ki se uporabljajo v korpusu:

```
<tagsdecl>
 <tagusage gi=text occurs=15>Element 'Text'.
  Attributes are LANG and ID.</tagusage>
 <tagusage gi=body occurs=15>Element 'Body'.
  Content model: TU+</tagusage>
 <tagusage gi=tu occurs=31900>Element 'Translation unit'.
  Attributes are LANG and ID.</tagusage>
 <tagusage gi=seg occurs=63800>Element 'Translation segment'.
  Attributes are LANG.</tagusage>
 <tagusage gi=s occurs=13386>Element 'Sentence'; only in 'orwl' text.
  Attributes are ID (value identical to original MTE bundle.</tagusage>
 <tagusage gi=w occurs=1091745>Element 'Word'.
  Attributes are TYPE (IMPLIED/971313, dig/26818, abbr/2662, comp/179)
  and (only in 'orwl' text) LEMMA, FUNCTION.</tagusage>
 <tagusage gi=c occurs=167243>Element 'Punctuation'.
  Attributes are TYPE (IMPLIED/131019, open/18115, close/18109).</tagusage>
 </tagsdecl>
```

Izjava o odgovornosti iz ene od glav besedil:

```
<respstmt>
<name>Jasna Belc, SVEZ</name>
 <resp lang="sl">Zagotovitev digitalnega originala</resp>
 <resp lang="en">Provision of digital original</resp>
 <name>&Scaron;pela Vintar, FF</name>
 <resp lang="sl">Poravnava</resp>
```

```
 <resp lang="en">Alignment</resp>
 <name>Toma&zcaron; Erjavec, IJS</name>
 <resp lang="sl">Tokenizacija, pretvorba v TEI</resp>
 <resp lang="en">Tokenisation, conversion to TEI</resp>
</respstmt>
```

Bibliografija izvornega besedila v glavi besedila:

```
<sourcedesc>
 <listbibl>
 <bibl lang="en" default="yes">
 <title lang="en">Linux Installation and Getting Started</title>
  <xref type="URL">http://metalab.unc.edu/LDP/LDP/gs/gs.html</xref>
  <xref type="URL">ftp://metalab.unc.edu/pub/Linux/docs/linux-doc-project/install-guide/</xref>
  <publisher>Specialized Systems Consultants
   <xref type="URL">http://www.ssc.com/</xref>
  </publisher>
 </bibl>
 <bibl lang="sl" default="no">
  <title lang="sl">Namestitev in za&ccaron;etek dela z Linuxom</title>
  <xref type="URL">http://www.lugos.si/delo/slo/LIGS-sl/</xref>
  <xref type="URL">ftp://ftp.lugos.si/pub/lugos/doc/install-guide-sl/</xref>
  <publisher>LUGOS: Linux User Group Of Slovenia
   <xref type="URL">http://www.lugos.si/</xref>
  </publisher>
 </bibl>
 </listbibl>
</sourcedesc>
```

Vsako dvobesedilo (element <body>) je sestavljeno iz prevodnih enot <tu>, od katerih vsaka vsebuje po dva segmenta <seg>: izvirnik in prevod. Definicija segmentov je del modula za osnovno jezikoslovno analizo TEI.analysis in lahko vsebuje raznovrstne oznake. Na{ korpus trenutno ozna~i besede, <w>, in lo~ila, <c>. Spodaj podamo nekaj prevodnih enot iz korpusa:

```
<tu lang="sl-en" id="kuca.303">
<seg lang="sl"><w>V</w> <w>takih</w> <w>krajih</w> <w>vsak</w><c>,</c> <w>mo&scaron;ki</w><c>,</c>
<w>&zcaron;enska</w> <w>in</w> <w>otrok</w> <w>i&scaron;&ccaron;e</w> <w>enako</w>
<w>pravico</w><c>,</c> <w>enako</w> <w>prilo&zcaron;nost</w><c>,</c> <w>enako</w>
<w>dostojanstvo</w><c>,</c> <w>brez</w> <w>diskriminacije</w><c>.</c></seg>
<seg lang="en"><w>Such</w> <w>are</w> <w>the</w> <w>places</w> <w>where</w> <w>every</w>
<w>man</w><c>,</c> <w>woman</w><c>,</c> <w>and</w> <w>child</w> <w>seeks</w> <w>equal</w>
<w>justice</w><c>,</c> <w>equal</w> <w>opportunity</w><c>,</c> <w>equal</w> <w>dignity</w>
<w>without</w> <w>discrimination</w><c>.</c></seg>
</tu>
...
<tu lang="sl-en" id="anx2.303">
<seg lang="sl"><w>Lahko</w> <w>se</w> <w>uporabljajo</w> <w>materiali</w> <w>iz</w> <w
type=abbr>tar.</w> <w type=abbr>&scaron;t.</w> <w type=dig>3003</w> <w>ali</w> <w type=dig>3004</w>
<w>pod</w> <w>pogojem</w><c>,</c> <w>da</w> <w>njihova</w> <w>skupna</w> <w>vrednost</w> <w>ne</w>
<w>presega</w> <w type=dig>20%</w> <w>cene</w> <w>izdelka</w> <w>franko</w> <w>tovarna</w><c>,</c>
<w>in</w></seg>
<seg lang="en"><w>However</w><c>,</c> <w>materials</w> <w>of</w> <w>heading</w> <w>No</w> <w
type=dig>3003</w> <w>or</w> <w type=dig>3004</w> <w>may</w> <w>be</w> <w>used</w> <w>provided</w>
<w>their</w> <w>value</w><c>,</c> <w>taken</w> <w>together</w><c>,</c> <w>does</w> <w>not</w>
<w>exceed</w> <w type=dig>20%</w> <w>of</w> <w>the</w> <w>ex-works</w> <w>price</w> <w>of</w>
<w>the</w> <w>product</w><c>;</c></seg>
</tu>
...
<tu lang="en-sl" id="ligs.303">
<seg lang="en"><w>Another</w> <w>modern</w> <w>text</w> <w>processing</w> <w>system</w> <w>is</w>
<w>TeX</w><c>,</c> <w>developed</w> <w>by</w> <w>Donald</w> <w>Knuth</w> <w>of</w> <w>computer</w>
<w>science</w> <w>fame</w><c>.</c></seg>
<seg lang="sl"><w>Drug</w> <w>sodoben</w> <w>sistem</w> <w>za</w> <w>stavljenje</w> <w>besedil</w>
<w>je</w> <w>TeX</w><c>,</c> <w>ki</w> <w>ga</w> <w>je</w> <w>razvil</w> <w>Donald</w>
<w>Knuth</w><c>,</c> <w>znan</w> <w>s</w> <w>podro&ccaron;ja</w> <w>teoreti&ccaron;nega</w>
<w>ra&ccaron;unalni&scaron;tva</w><c>.</c></seg>
</tu>
```

Oznake na ravni besed seveda niso mi{ljene za branje, pa~ pa, da olaj{ajo nadaljnje ra~unalni{ko izkori{~anje korpusa. Kot se vidi tudi iz zgornjih primerov, smo besednim oznakam dodelili tudi nekaj tipov, ki utegnejo biti koristni za nadaljne obdelave:

– <w type=comp> Ve~besedna leksikalna enota, npr. *medtem ko, vice versa, New York*
– <w type=dig> Beseda, ki vsebuje {tevilke, npr. *1984, 3., IV, 20%, 1993-1996, 25/76, 16MB*
– <w type=abbr> Kratica (beseda, ki se kon~a s piko), npr. *tar., et al., S.u.S.E., dipl.*
– <w> privzeti tip za "navadne" besede, npr. Slovenije, market, 's, Article, 'ivinorejo, INAVGURACIJSKI, Hurt-Andreatta, Hrup51, E-po{tni, D'you

^eprav je korpus ozna~en po TEI berljiv s poljubnim urejevalnikom besedil, uporabniku ni prav prijazen. Ena od odlik zapisa SGML naj bi bila enostavna pretvorba v format, ki je primeren konkretni aplikaciji. Za bolj{o preglednost smo zato naredili pretvorbo iz zapisa TEI v zapis HTML, tako da so vse glave korpusa ter primeri besedil na voljo preko standardnih mre'nih brkljalnikov.

Dostopnost

Doma~a stran IJS-ELAN ima naslov http://nl.ijs.si/elan/. Poleg informacij o korpusu (npr. glave TEI v zapisu HTML) je tam dostopen tudi celoten korpus v zapisu TEI. Glave korpusa in posameznih delov dolo~ajo, da je korpus prosto dostopen pod pogojem, da se citira njegove vire, dokumentirane v glavi.

Za uporabo besedil ozna~enih po TEI je potrebno vsaj osnovno znanje programiranja in dostop do orodij za izkori{~anje korpusa. Da bi olaj{ali uporabo korpusa, omogo~amo tudi mre'ne konkordance nad korpusom. Konkordance so eden osnovnih na~inov uporabe korpusov, ki omogo~ajo iskanje in izpis besed in besednih zvez skupaj s sobesedilom. Na IJS imamo licenco za uporabo konkordan~nika CQP (Christ 1994), ki je hiter, ima bogat iskalni jezik ter podpira prikaz poravnanih segmentov. Na Univerzi v Gothenburgu so razvili spletni vmesnik za CQP, ki smo ga nato izbolj{ali in priredili za delo s slovenskim naborom znakov ter za prikaz poravnanih segmentov. Korpus IJS-ELAN smo pretvorili v format, ki ga zahteva CQP in ga vklju~ili v nabor preko mre'e dostopnih korpusov.

S konkordan~nikom, ki ga najdemo na http://nl2.ijs.si/corpus/index-bi.html, je mogo~e iskati po angle{kem ali slovenskem delu korpusa, pri ~emer lahko dodatno podamo omejitev glede na poravnani segment. Tako lahko npr. i{~emo vse pojavitve besede 'drevo', kjer se v prevedenem segmentu ne pojavi beseda 'tree'. Iskalni izrazi so lahko enostavni (npr. 'corpus' ali 'besed*') ali pa uporabljajo polno (in razmeroma kompleksno) sintakso CQP. Slednja dopu{~a poljubne regularne izraze nad nizi, iskanje po med seboj oddaljenih delih segmentov ter iskanje po tipih besed oz. lo~il. Mre'ni vmesnik ponuja tri na~ine izpisovanja zadetkov: privzeti na~in je dvojezi~ni, kjer zadetku s sobesedilom sledi tudi poravnani segment v drugem jeziku; enojezi~ni izpis dobimo v na~inu KWIC (key-word in context) ali pa kot frekven~ni seznam zadetkov.

Zaklju~ki

V ~lanku smo predstavili slovensko-angle{ki vzporedni korpus IJS-ELAN. Korpus je uporaben kot vir prevodov terminov, predvsem skozi mo'nost mre'nih konkordanc, pa tudi kot podatkovna zbirka za raziskave in razvoj jezikovnih tehnologij, predvsem tistih vezanih na terminologijo in prevajanje. Predstavljeni korpus je prosto dostopen v upanju, da dodatno spodbudi razvoj korpusnega jezikoslovja za slovenski jezik. ^eprav je za slovenski jezik 'e na voljo oziroma v delu nekaj korpusov, npr. korpus FIDA (Krek et al.), so le ti enojezi~ni; IJS-ELAN tako predstavlja prvi ve~ji dvojezi~ni korpus za na{ jezik.

Nadaljnje delo s korpusom bo usmerjeno predvsem v zagotovitev bogatej{ega nabora oznak. Tu je na prvem mestu lematizacija in oblikoslovno ozna~evanje besed v korpusu. V

dalj{i perspektivi pa bi bilo seveda najbolj koristno pove~ati koli~ino besedil v korpusu, kar pa bi bilo mo‘no samo v okviru {ir{ega projekta.

Zahvale

Pri delu na korpusu, predstavljenem v tem ~lanku, so sodelovali Roman Maurer, Andrej Skubic in [pela Vintar. Besedila za korpus so prispevali uradi in slu‘be Republike Slovenije, posebej {e Slu‘ba Vlade RS za evropske zadeve. Besedila so tudi prispevali Linux Users Group of Slovenia, LUGOS in Lek d.d., OTC Division. Delo na projektu je deloma financirala pogodba z Institut fuer deutsche Sprache v okviru projekta MLIS-ELAN 121 in pogodba MZT L2-0461-0106 .

LITERATURA

Susan ARMSTRONG, Masja KEMPEN, David MCKELVIE, Dominic PETITPIERRE, Reinhardt RAPP, Henry THOMPSON, 1998: Multilingual corpora for cooperation. *Proceedings of the First International Conference on Language Resources and Evaluation*. LREC’98. Granada: ELRA. 579–980.

Oliver CHRIST, 1994: A Modular and Flexible Architecture for an Integrated Corpus Query System. *Proceedings of COMPLEX ’94: 3rd conference on Computational Lexicography and Text Research*. Budimpe{ta. URL: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Philippe di CRISTO, 1996: *Mtseg: The multext multilingual segmenter tools*. MULTEXT Deliverable MSG 1. Version 1.3.1. CNRS. Aix-en-Provence. URL: http://www.lpl.univ-aix.fr/projects/multext/MtSeg/

Ludmila DIMITROVA, Toma‘ ERJAVEC, Nancy IDE, Heiki-Jan KAALEP, Vladimir PETKEVI^, Dan TUFIS, 1998: Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *COLING-ACL ’98*. Montreal, Quebec. 315–319.

Toma‘ ERJAVEC, 1999a: A TEI encoding of aligned corpora as translation memories. *Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99)*. Bergen: ACL.

– – 1999b: Making the ELAN Slovene/English Corpus. *Proceedings of the Workshop on Language Technologies – Multilingual Aspects*. Ljubljana: Univerza v Ljubljani. 23–30.

Toma‘ ERJAVEC, Nancy IDE, 1998: The MULTEXT-East corpus. *First International Conference on Language Resources and Evaluation, LREC’98*. Granada: ELRA. 971–974.

Toma‘ ERJAVEC, Ann LAWSON, Laurent ROMARY (ur.), 1998. *East meets West: A Compendium of Multilingual Resources*. CD-ROM, TELRI Association e.V. URL: http://www.ids-mannheim.de/telri/cdrom.html

Stig JOHANSSON, Jarle EBELING, Knut HOFLAND, 1996. Coding and aligning the English-Norwegian parallel corpus. *Languages in Contrast*. Ur. K. Aijmer, B. Altenberg, M. Johansson. Loud: Lund University Press. 87–112. URL: http://www.hit.uib.no/enpc/

Simon KREK, Marko STABEJ, Vojko GORJANC, Toma‘ ERJAVEC, Miro ROMIH, Peter HOLOZAN. *FIDA: korpus slovenskega jezika*. URL: http://www.fida.net

Tony MCENERY, Andrew WILSON, Fernando SANCHEZ-LEON, Amalio NIETO-SERRANO, 1997: Multilingual Resources in European Languages: Contributions of the CRATER Project. *Literary and Linguistic Computing* 12/4.

C. M. SPERBERG-MCQUEEN, Lou BURNARD (ur.), 1994: *Guidelines for Electronic Text Encoding and Interchange*. Chicago, Oxford. URL: http://www-tei.uic.edu/orgs/tei/

[pela VINTAR, 1999: A Lexical Analysis of the ELAN Slovene-English Corpus. *Proceedings of the Workshop on Language Technologies – Multilingual Aspects*. Ljubljana: Univerza v Ljubljani. 63–70.

Toma‘ Erjavec
Institut Jo‘ef Stefan, Ljubljana
Odsek za inteligentne sisteme